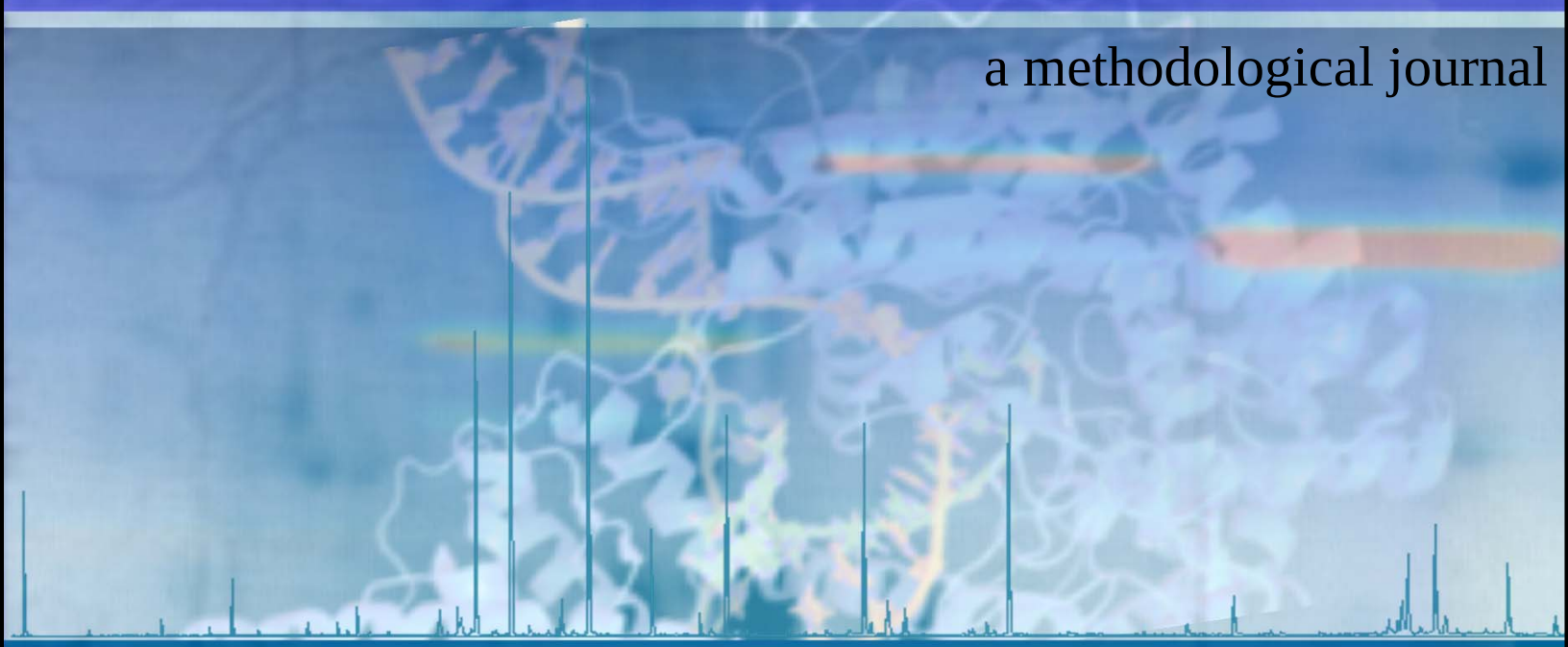
Volume 1 Issue 1 1 February 2011 ISSN 2182-0287

Journal of Integrated OMICS

a methodological journal



JIOMICS

Journal of Integrated OMICS

Focus and Scope

Journal of Integrated OMICS, JIOMICS, provides a forum for the publication of original research papers, preliminary communications, technical notes and critical reviews in all branches of pure and applied "-omics", such as genomics, proteomics, lipidomics, metabolomics or metallomics. The manuscripts must address methodological development. Contributions are evaluated based on established guidelines, including the fundamental nature of the study, scientific novelty, and substantial improvement or advantage over existing technology or method. Original research papers on fundamental studies, and novel sensor and instrumentation development, are especially encouraged. It is expected that improvements will also be demonstrated within the context of (or with regard to) a specific biological question; ability to promote the analysis of molecular mechanisms is of particular interest. Novel or improved applications in areas such as clinical, medicinal and biological chemistry, environmental analysis, pharmacology and materials science and engineering are welcome.

Editors-in-Chief

Associated editors

Editorial

Journal of Integrated OMICS – Where no journal has gone before

The first International Congress on Analytical Proteomics – ICAP - was held in Caparica, Portugal, in 2008. This congress was conceived to cover a gap in the proteomics world, namely a direct link to the rigorous techniques of analytical chemistry through method development. Other international congresses such as HUPO or EuPa cover method development but not in a monographic format. During this first ICAP, discussions took place among the participants regarding the opportunity of promoting a new forum for discussion within the galaxy of journals devoted to proteomics. It was soon agreed that the new journal should meet some specific criteria that would clearly establish the technical interface between 'Omics approaches and analytical chemistry that was to be supported and promoted. Thus, it was established that a journal of the 21st century should not only be open access and thus fully available to everyone, but also open in many different ways. This operating philosophy thus defined a set of novel characteristics for the nascent journal. First, as has been said, access should it be open to all researchers. Second, the target areas of the new journal should cover the breadth of disciplines defining the 'Omics arena, not only proteomics. Third, publication of papers dealing with methodological development should be encouraged, but the journal should not be restricted only to this. Finally, and uniquely, the Journal should be published by its own staff. Accordingly, a low fee of about 100 Eu was set as the regular price to be paid for publishing in the Journal. After extensive debate, the name of the new journal was set as Journal of Integrated OMICS — A Methodological Journal. We thank Professor Jens Coorssen, (University of Western Sydney, Australia), as he first suggested this name that finally became the official title of the new Journal. It was also decided to open the Editorial Board to all researchers with demonstrated experience in any OMIC field who were interested to help in any way to develop this new initiative — to serve as editor, referee, or both.

The next step in establishing the new Journal was to create an official site. Professor Florentino Fdez-Riverola (University of Vigo, Spain), contributed substantially to this task, and the website was launched April 9, 2010. The excellent result can viewed via the link www.jiomics.com. The website also includes a Forum, http://www.bioscopegroup.org/jiomicsforum/, recently created to support collegial debate on subjects of interest for the OMICS community such as new advances in mass spectrometry or chromatography.

Last, but not least, it was necessary to communicate the existence of this new International Journal to the scientific community, requesting the cooperation of all those who would like to collaborate with this initiative. Therefore an international call was made requesting members for the Editorial Board; this was a great success and to date, there are more than 200 associated editors. We are proud to say that our membership reflects true international breadth.

The ISSN of the Journal was obtained in January 2011. IDOI will be assigned to our academic papers, and we hope to also be indexed in SCOPUS and SciFinder in the near future.

We certainly wish to acknowledge all authors participating in the launching of the first issue. On the horizon, other researchers will consider JIOMICS as their regular publishing media. For their contributions, we are indebted to Gilberto Igrejas, Qing-Yu He, Alan J. Tacketta, Marco A. Z. Arruda, Guoping Zhang, Chantagan Srisomsap, Randen L. Patterson, Damian B. van Rossum, Ganesh Chandra Sahoo, Elena López, Rune Matthiesen, Jan-Jaap Wesselink, Antonio Ferreira, Kay Ohlendieck, Bernard Corfe, Janina M. Tomm, Ying-Ray Lee, Hong-Lin Chan, Maria Moreno, Pieter de Lange, Eiji Kinoshita, Vladimir Lazarevic, and Martin von Bergen. Thanks indeed for entrusting us with your work.

Next July, 2011, the 2nd ICAP (http://sing.ei.uvigo.es/ICAP/) will be held in Ourense, Spain. A number of issues regarding the journal will be discussed there, including the option to open for general comment the reviewing of the papers and other questions that any member of the editorial board might raise.

We would like to close by encouraging all the editorial staff to contribute actively to the success of JIOMICS, because this journal belongs to all of us.

Ourense, February 2011

On behalf of the Editors-in-Chief,

José-Luis Capelo-Martínez

Department of Physical Chemistry. Faculty of Sciences of Ourense. University of Vigo at Ourense Campus. 32004-Ourense, Spain. E-mail address: jlcapelom@uvigo.es

JOURNAL OF INTEGRATED OMICS

A methodological Journal

CONTENTS OF VOLUME 1 | ISSUE 1 | FEBRUARY 2011

REVIEW ARTICLES

ORIGINAL ARTICLES

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

REVIEW ARTICLE | DOI: 10.5584/jiomics.v1i1.40

Functional phosphoproteomics tools for current immunological disorders research

Elena López*[2,θ], Rune Matthiesen[3], Isabel López[4], Keith Ashman[2], Jesús Mendieta[5,6], Jan-Jaap Wesselink[6], Paulino Gómez-Puertas[5], Antonio Ferreira*[1].

[1]Immunology Unit, Hospital Universitario La Paz, Pº de la Castellana, 261 28046 Madrid, Spain; [2]Biotechnology Programme, Spanish National Cancer Research Centre (CNIO), C/ Melchor Fernández Almagro, 3, 28029, Madrid, Spain; [3]Institute of Molecular Pathology and Immunology of the University of Porto, Rua Dr. Roberto Frias s/n 4200-465 Porto, Portugal; [4]Department of Hematology, Hospital Universitario 12 Octubre, Avda de Córdoba s/n Madrid, 28041, Spain; [5]Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM) Campus de Cantoblanco, c/Nicolás Cabrera, 1, 28049 Madrid, Spain; [6]Biomol–Informatics, S.L., Parque Científico de Madrid, Campus de Cantoblanco, c/Faraday 7, 28049 Madrid, Spain; [θ]Current address: Centro de Investigación del Hospital Universitario 12 de Octubre, Avda de Córdoba s/n Madrid, 28041, Spain.

Received: 24 September 2010 **Accepted:** 13 October 2010 **Available Online**: 15 October 2010

ABSTRACT

Signaling networks are key elements in all major aspects of cellular life, playing a major role in inter- and intracellular communications. They are involved in diverse processes such as cell-cycle progression, cellular metabolism, cell-cell communication and appropriate response to the cellular environment. The latter comprises a whole range of networks that are involved in regulation of cell development, differentiation, proliferation, apoptosis, and immunologic responses. The key mechanism involves the transduction of extracellular signals across the cell surface to different effectors in the cytosol and the nucleus. Dysregulation of these pathways is often associated with immunology disorders and malignant diseases such as cancer. One of the most common mechanisms of activation and/or inactivation of signaling transduction pathways is phosphorylation and de-phosphorylation at serine, threonine and tyrosine residues. Phosphoproteomics is playing an important role in our understanding of how phosphorylation participates in translating distinct signals into the normal and or abnormal physiological responses, and has shifted research towards screening for potential therapies for diseases and in-depth analysis of phosphoproteomes. Given the importance of phosphoproteomics in translational research we aim at outlining phosphoproteomic approaches based on mass spectrometry (MS). This review focuses on (1b) the role of phospho signaling in immunology, (2a) current phosphopeptide enrichment methods based on IMAC and titanium dioxide, (2b) phosphopeptide analysis by MS, and (2c) issues concerned with interpretation of phospho spectra by database dependent search. Finally, quantitative methods used in phosphoproteomics such as Stable Isotope labeling with Amino acid in cell Culture (SILAC), isobaric Tag for Relative and Absolute Quantitation (iTRAQ) and Absolute Quantification (AQUA) is discussed in section 3.

Keywords: Phosphorylation; Proteomics; Mass Spectrometry; Immunology disorders.

Abbreviations

AQUA Absolute Quantitation; **CID** Collision-Induced Dissociation; **Da** Dalton (molecular mass); **ECD** Electron Capture Dissociation; **ESI** Electron Spray Ionization; **ETD** Electron Transfer Dissociation; **FT-ICR** Fourier transform-Ion Cyclotron Resonance; **H_3PO_4** Phosphoric acid; **ICR** Ion Cyclotron Resonance; **IMAC** Immobilized Metal Affinity Capture; **IT** Ion Trap; **iTRAQ** isobaric Tag for Relative and Absolute Quantitation; **kDa** kilodalton (molecular mass); **LC** Liquid Chromatography; **MALDI** Matrix-Assisted Laser Desorption/Ionization ; **MOAC** Metal Oxide Affinity Chromatography; **Mr** Relative molecular mass (dimensionless); **MS** Mass Spectrometry; **MS/MS** tandem mass spectrometry; **m/z** Mass to charge ratio; **PID** Primary Immuno deficiencies; **PTM** Post-Translational Modification; **SILAC** Stable Isotope Labelling with Amino acid in cell Culture; **SIMAC** Sequential Elution from IMAC; **TiO_2** Titanium dioxide; **TOF** Time Of Flight.

***Corresponding author:** Elena López. Email Adress: elopezv@h12o.es; Antonio Ferreira. Email Adress: aferreira.hulp@salud.madrid.org.

Index

1. Introduction

1.1 *A short history of phosphorylation studies.*

In 1954 the first protein kinase activity was observed by using ^{32}P labeling [1]. The next major finding occurred in 1979, when Tony Hunter and co-workers discovered the first tyrosine phosphorylated by a signaling cascade [2]. This was shortly followed by the elucidation of the amino acid sequence of the first kinase (PKA) by Edman degradation [3]. The spotlight in the late eighties and nineties was focused on the characterization of signaling cascades, mainly the MAP kinases [4]. Indeed, this period witnessed the purification and identification of the first protein tyrosine phosphatase [5]. One of the major landmarks of the late 1980s was the creation of the first protein kinase knockout in mouse [6]. In 1997, the use of sophisticated genetic engineering allowed the manipulation of the amino acid residues in tyrosine kinases [7]. Recently, Ficarro *et al.* [8] and Gruhler *et al.* [9] reported the identification of thousands of phosphopeptides including the phosphorylation sites from yeast by using a combination of phosphoproteomic tools such as phosphopeptide enrichment and tandem MS. In 2002 the proportion of the human genome encoding protein kinases was analyzed and for the first time a complete set of human kinase sequences were assembled and classified [10]. This study defined the human kinome, which contains around 500 protein kinases, many of which are considered likely therapeutic targets. In recent years large scale phosphoproteome analysis quantitatively profiled thousands of phosphopeptides [11, 12, 13, 14, 15, 16, 17, 18]. Finally, bioinformatics has become an important team player when the phosphopeptide results from latest generation of mass spectrometers needs to be tested for association with existing molecular information (Figure 1).

1.2. *Relevance of phosphorylation in immunology disorders*

During an infection caused by agents like fungi, bacteria, or viruses, pathogen components bind to receptors on immune cells, and trigger the activation of signaling pathways [19, 20, 21]. This activation stimulates the production of inflammatory mediators, including pro-inflammatory cytokines like tumor necrosis factor, interleukin-1, chemokines, prostaglandins and interferons. Interestingly, the precise production of inflammation mediators depends on body site and type of inflammation. All these mediator components can act locally and are also secreted into the blood where they mount responses to fight the invading pathogen. However, the uncontrolled production of these compounds can cause tissue

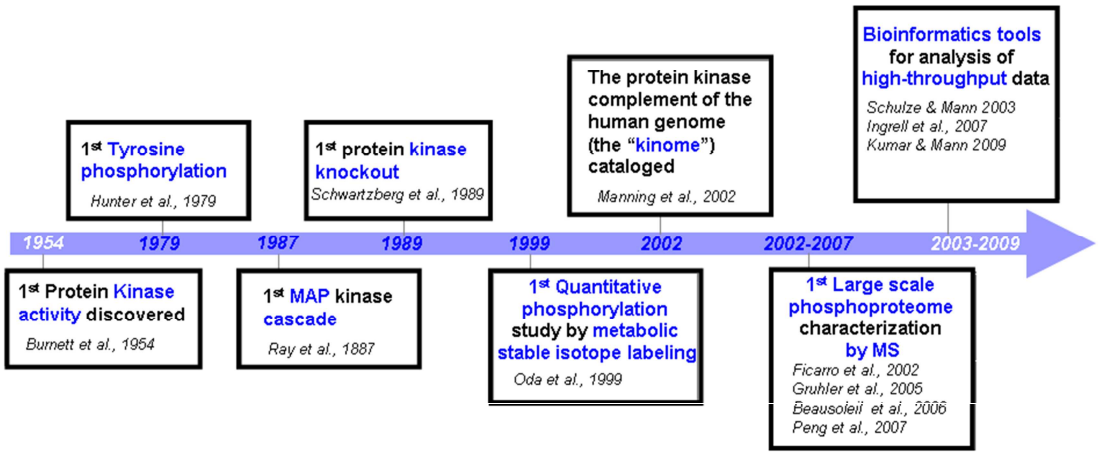


Figure 1. Evolution of protein phosphorylation studies in phosphoproteomics. Different studies on protein-phosphorylation and recent progress on MS based proteomics are shown for the last fifty years. In the beginning analysis of specific kinases by classical molecular biology techniques initiated the interest in phosphoproteomics. Nowadays, large-scale studies can be carried out thanks to advances on MS-based proteomics tools and computational methods.

damage. Increased levels of pro-inflammatory cytokines are also characteristic of chronic inflammatory and autoimmune diseases, like for example: Polymyalgia Rheumatica, Addison's disease, Systemic lupus erythematosus, Type 1 Diabetes Mellitus, inflammatory bowel disease, asthma, rheumatoid arthritis and psoriasis [22, 23, 24].

The most commonly used anti-inflammatory drugs are glucocorticoids but there are side effects associated with the long-term use of steroids. There is a need for anti-inflammatory drugs with fewer side effects, orally active, more effective and less expensive. The potential targets of such drugs could be components of the signaling pathways that regulate the production of inflammatory mediators.

While autoimmune diseases are complex and involve de-regulation of both the innate and adaptive immune system. The study of the response of innate immune cells to the components of the pathogens which bind the immune receptors proved to be a productive way of dissecting the signaling pathways which control the production of pro-inflammatory cytokines.

Recently, protein-kinases have become the pharmaceutical industry's most studied class of drug targets and a number of drugs that inhibit these enzymes have been approved for the treatment of cancers. Additionally, recent evidence suggests that several protein kinases of the innate immune system are potentially attractive targets for the development of drugs to treat chronic inflammatory diseases [25].

Cancer, at a molecular level, is considered a state of altered signaling. One of the most common mechanisms of activation and/or inactivation signaling pathways is phosphorylation and de-phosphorylation at serine, threonine and tyrosine residues. This type of modification controls a variety of cellular processes including cellular growth, proliferation, cell cycle control, cytoskeletal mobility and receptor regulation [26]. Phosphorylation leads to allosteric modifications that may result in sufficient conformational changes that cause activation or inactivation of various proteins and associated altered functioning. We hypothesize that identification of phosphoproteins associated with the various stages of cancer may provide information on the mechanism of tumorigenesis and insights into the development of diagnostic and therapeutic procedures. The mitogen activated protein kinase (MAPK) pathways are known to be deregulated in many human malignancies [27, 28, 29, 30]. The best studied MAPKs, with regards to malignancy, are the extracellular signal regulated protein kinases (ERK). ERKs phosphorylate cytoplasmic targets or migrate to the nucleus where they can activate transcription factors involved in cellular proliferation. Aberrant signaling in the MAPK/ERK pathways has been described in prostate, breast and colon cancers *in vitro* as well as *in vivo* models [31, 32, 33, 34]. In addition, in cervical cancer, one study has described decreased activation of ERK1/2 in invasive cervical carcinoma [35]. A second relevant example is annexin A1, which is a calcium dependent phospholipid binding protein that has been linked to membrane trafficking through exocytosis and endocytosis [36].

Other studies have evaluated the role of annexin A1 in the modulation of the MAPK/ERK [37]. Indeed, many members of the Annexin family are known to undergo alternative splicing yielding a number of isoforms. The resultant variant forms may have different functions and binding capacity compared to the native forms [38]. Another relevant example is related to the DNA-Protein Kinase catalytic subunit (DNA-PKcs), a macromolecule found to be involved in the repair of double stranded DNA breaks through activation of p53 which was found to be expressed in cancer specimens in its tyrosine phosphorylated and cleaved form [39]. In contrast, in normal specimens DNA-PKcs existed in its intact, full length and non-phosphorylated form. The aim of this study was to identify differential expression and modification of proteins that could suggest aberrant pathways which could serve as novel targets for developing new therapies in the treatment of cervical cancer and aide in monitoring disease recurrence or progression.

In addition, we will provide examples of the application of phosphoproteomics in immunological studies. The CDC25 family of proteins is comprised of dual specificity phosphatases that regulate cell cycle transitions, and are key targets of the checkpoint machinery to maintain genome stability during DNA damage. Three isoforms of CDC25 have been identified in mammalian cells: CDC25A, CDC25B, and CDC25C. CDC25A and CDC25B overexpression has been reported in many types of human cancers, but is insufficient to cause cancer, and the mechanism responsible for CDC25 overexpression is unclear [40, 41].

Studying dose-response effects of the anti-cancer drug rapamycin on the phosphoproteomics level has identified hundreds of novel rapamycin-targeted cellular proteins and their phosphorylation sites. This information enabled us to identify CDC25B as the key enzyme in mediating rapamycin induced oncogenic AKT activation. Importantly, we show that phosphoproteomic profiling of a certain therapeutic agent cannot only identify potential drug target(s) to improve the efficacy of that therapeutic approach in disease treatment, but can also provide cellular information on possible beneficial and adverse side effects of a certain disease therapy when treating patients [42].

Primary Immunodeficiencies (PID) are "nature's experiments" which have allowed not only the elucidation of many signaling pathways but also their function and clinical relevance.

An interesting example is related to Bruton's tyrosine kinase (Btk; member of the Tec family of kinases) [43, 44], important in B-lymphocyte development, differentiation, and signaling. Btk is predominantly expressed in B lymphocytes and monocytes but not in plasma cells [45, 46]. Btk expression in the B-cell lineage is also developmentally regulated, with bone marrow-derived hematopoietic stem cells, common lymphoid progenitor cells, developing B and myeloid lineages showing the highest levels, whereas resting mature cells prior to activation have reduced cellular Btk. As B lymphocytes are the only cells known to be affected in X-linked

agammaglobulinemia (XLA), the physiological significance of Btk expression in other cell types remains to be established.

Mutations in the Btk gene lead to XLA in humans and X-linked immunodeficiency (Xid) in mice. Activation of Btk triggers a cascade of signaling-events that culminates in the generation of calcium mobilization and fluxes, cytoskeletal rearrangements, and transcriptional regulation involving nuclear factor-κB (NF-κB) and nuclear factor of activated T cells (NFAT). In B cells, NF-κB was shown to bind to the Btk promoter and induce transcription, whereas the B-cell receptor dependent NF-κB signaling pathway requires functional Btk. Moreover, Btk activation is tightly regulated by a plethora of other signaling proteins including protein kinase C (PKC), Sab/SH3BP5, and caveolin-1. Additionally, the prolyl isomerase Pin1 negatively regulates Btk by decreasing tyrosine phosphorylation and steady state levels of Btk [47]. It is intriguing that PKC and Pin1, both of which are negative regulators of Btk, bind to the pleckstrin homology domain of Btk. To this end, novel mutations in the pleckstrin homology are being investigated, in order to design selective and novel drugs [48].

Another relevant example is related to IgA deficiency [49, 50, 51]. IgA plays a role in systemic immunity which is still not understood, IgA antibodies in secretions can neutralize viruses, bind toxins, agglutinate bacteria, prevent bacteria from binding to mucosal epithelial cells, and bind to various food antigens preventing entry into the general circulation (Table 1).

Table 1. Examples of specific human IgA antibody reactivity

(a) Bacteria	Bacteroides fragillis
	Bordetella Pertussis
	Campylobacter
	Clostridiumdiphtheriae, C. tetani
	Escherichia coli O, K antigens, enterotoxin
	Salmonella
	Shingella
	Streptococcus
	Streptococcus mutans
	Neisseria gonorrhoeae
	Vibrio cholerae
(b) Fungi	Candida albicans
(c) Protozoa	Giardia
(d) Viruses	Arboviruses - Semliki forest
	Coxsackie virus
	Cytomegalovirus
	Dengue
	Echovirus
	HIV
	Influenza A
	Parainfluenza
	Poliovirus 1, 2, 3
	Respiratory syncytial virus
	Rhinovirus
	Rotavirus
	Ross river
	Japanese B
(e) Others	Milk proteins
	Peanut lectin
	Soy lectin
	Wheat gluten, gliadin

Because IgA is known to protect mucous surfaces, it is a mystery why most IgA-deficient subjects remain healthy. This lack of disease in IgA deficiency is often attributed to a compensatory increase in IgM in the secretions [52]; the colostrum of IgA-deficient subjects has been shown to contain abundant amounts of IgM. The main immunological difference between the IgA-deficient and the normal intestinal tract is the substitution of IgM secreting plasma cells for IgA-secreting cells [53]. This difference is evident in both healthy and ill IgA-deficient subjects.

When nodular lymphoid hyperplasia develops, the nodules contain a proliferation of IgM plasma cells.

Despite the fact that most IgA-deficient subjects are not ill, IgA deficiency has been associated with an astonishing number of specific disorders (Table 2) [51, 54, 55, 56].

A fundamental defect in IgA deficiency is the failure of IgA-bearing B lymphocytes to mature into IgA secreting plasma cells. This appears to be a defect of stem cells, since IgA deficiency can be transferred by bone marrow engraftment. A cardinal feature is that there is a paucity of IgA-bearing gastrointestinal plasma cells. There are decreased (but not absent) numbers of IgA-bearing B cells in the peripheral circulation in these patients, which bear an immature phenotype; that is, IgA-bearing B cells that also are positive for IgM and IgD. The reason that IgA deficient subjects have B cells that fail to switch to the production of IgA is unknown. There is much to learn about the role of IgA in human immunity. IgA is the most abundant immunoglobulin made and also the most ignored. IgA is the second-most prevalent immunoglobulin in serum, but its role in systemic immunity is unknown. IgA has a known biological activity against a number of pathogens and can be shown to serve as a barrier preventing the permeation of foreign antigens and pathogens; on the other hand, lack of IgA does not usually result in a perceptible immune defect [57].

Because no genetic defect is known in IgA deficiency and its clinic diversity, many genetic studies have been focused on the relationship of IgA deficiency with HLA system [58, 59, 60, 61, 62, 63, 64], complement components [65, 66], transmembrane activator and calcium modulator and cyclophilin ligand interactor (TACI) [67, 68, 69] and other genes [70, 71, 72, 73, 74]. Recently, a genome-wide association study identified association with Interferon-induced helicase C domain containing protein 1(IFIH1) and C-type lectin domain family 16 (CLEC16A) [75].

The phosphoproteomics of IgA deficiency could explain not only why IgA deficient people are asymptomatic but also why some patients present infections and/or autoimmunity

Table 2. Examples of conditions associated with selective IgA deficiency

(a) Allergy
Asthma, atopy, eczema
(b) Autoimmunity
Rheumatoid arthritis ITP, hemolytic anemia, pernicious anemia, systemic lupus erythematosus, Still's disease, transfusion reactions due to anti IgA antibody, dermatomyositis, vitiligo, Sjogren's syndrome,Henoch–Schonlein syndrome, primary biliary cirrhosis, autoimmune hepatitis
(c) Endocrinopathy Thyroiditis
Graves disease, idiopathic Addison's disease, diabetes mellitus, 21-hydroxylase deficiencycd
(d) Gastrointestinal diseases
Giardiasis, Crohn's disease, ulcerative colitis, nodular lymphoid hyerplasia, celiac disease, lactose intolerance, malabsorption, villous atrophy, achlorhydria, cholelithiasis
(e) Malignancy
Gastric carcinoma and lymphoma
(f) Neurological
Seizures, migraine, sensory neuropathy, myasthenia gravis, cerebral vasculitis
(g) Respiratory tract
Recurrent sinopulmonary infections, sarcoidosis, pulmonary hemosiderosis

and/or other pathological situations.

XLA and IgA deficiency are two relevant examples of PID because in XLA, the genetic disorder was observed for the first time in a PID and in IgA deficiency, the most frequent PID, the genetic disorder is unknown yet.

Many other PID are interesting for analysis with phosphoproteomics tools because their genetic defects involve many signaling pathways depending on phosphorylation/dephosphorylation phenomena, such as IL-1 receptor associated kinase 4 (IRAK 4) deficiency with functional defects of TIR-IRAK signaling pathway and bacterial infection (pyogenes) in the patients [76]. Another example is the Hyper IgE syndrome, a clinical situation where three different genetic disorders including protein phosphorylation can cause the disease, it has mutations in either signal transducer and activator of transcription protein 3 (STAT 3) [77, 78] or tyrosin kinase 2 (TYK 2) [79] or in the dedicator of cytokinesis 8 (DOCK 8) [80].

For up to date information on PID with defects in signaling pathways involving phosphorylation, see the reference of the International Union of Immunology Societies Expert Committee on Primary Immunodeficiencies [81].

Current phosphoproteomics strategies allow the study of different immunological diseases. The advantage over other methodologies, is that they allow the assignment of the specific amino acids which are phosphorylated and or dephosphorylated when comparing healthy and malignant tissues/cells. In addition, phosphoproteomics gives spatial and temporal biological information. Therefore, it will help to establish the right and more specific prognosis and diagnosis, and therefore to treat different immunological disorders with more accurate therapies. Indeed, these findings suggest that it plays a critical role in studies related to the discovery of therapeutic targets [26, 82].

Nowadays, many hospitals have included and are including MS technologies in order to collaborate with/and for clinical research, covering many kinds of pathologies. Thus, improving the therapy - treatments of the patients.

1.3. *The role of phosphorylation in protein structure and function*

Reversible phosphorylation of proteins modulates a great variety of cellular processes. Biological regulation and signaling is in this manner controlled by kinases, phosphatases and phosphopeptide recognition domains of effector proteins [83]. Peptides phosphorylated mainly on serine, threonine or tyrosine residues are specifically recognized by a variety of protein domains. Phosphotyrosine (pTyr)-containing peptides interact with SH2 (Src Homology 2) and PTB (Phosphotyrosine Binding) domains. Phosphoserine and phosphothreonine (pSer/pThr) are recognized by an increasing variety of modular protein domains, including 14-3-3, WW, FHA (Forkhead Associated), Polo-box, and BRCT (BRCA1 C-terminal) among others [83, 84]. In plants, only 14-3-3 and FHA domains are found to mediate phosphopeptide binding processes [85].

According to the classical theory, all these domains discriminate between the phosphorylated vs. the non-phosphorylated state of proteins, based on the conformational changes induced by the presence of a negatively-charged phosphate group in the basal state of the phosphopeptide. Structure/function experiments performed using NMR relaxation techniques on the phosphorylated state of the response regulator "Nitrogen regulatory protein C" from Salmonella typhimurium [86] indicated that both states (active and inactive) are present before phosphorylation, suggesting that this process only acts through the modification of the preexisting equilibrium. Changes induced by phosphorylation could be of different magnitude. In some cases, adding of phosphate groups leads to partially unfolding of previously structured domains, making them more accessible to phosphopeptide recognition domains, as it has been demonstrated by NMR studies in the case of the KH-domain 1 of the RNA binding K-homology splicing regulator protein and its interaction to 14-3-3 protein [87].

Since no high resolution 3D coordinates obtained by NMR or crystallization procedures are available, the increasing capacity and accuracy of computational simulations allow us to generate models of the different phosphorylation states of

a protein. As an example, these methods have been used to explore low-energy conformations of the phosphorylation-regulated structure of the R domain of the cystic fibrosis transmembrane conductance regulator [88]. The results suggested that the increment of the radius of gyration of the phosphorylated states of the domain was correlated with the activation of the chloride channel through reorganization of its helix packing.

Phosphorylation can not only be correlated with changes in tertiary structure but also with changes in secondary structure. A recent study [89], monitoring the secondary structure changes of a coiled-coil model peptide through Circular Dichroism, found a sequential cascade starting from a stable helical state continued by a phosphorylation-induced unfolded structure and finishing, in this particular case, in a magnesium stabilized alpha-helix structure. Computational simulation could also be very useful to model this type of changes. A good example of this could be the comparative study, using both Molecular Dynamics and Circular Dichroism techniques, of the displacement in the equilibrium between the unphosphorylated peptide of the third tau microtubule-binding repeat and its mono- and di-phosphorylated states [90]. Molecular Dynamics models predicted a stabilization of the peptide in a short alpha helical structure mediated by phosphorylation. This arrangement, contrasted using Circular Dichroism approaches, will lead the peptide to a homo-polymerized complex, connecting the structural changes due to phosphorylation with modifications in the functional behavior of the peptide (Figure 2).

In recent years, simulation techniques based on Quantum Mechanics are increasing their capacity and are being implemented at a scale that allows the simultaneous calculation of the behavior of several dozens of atoms. Particularly, these methods permit the analysis of the formation and breaking of covalent bonds in the active center of proteins (Quantum Mechanics) within the framework of the atomic models of complete biological macromolecules, calculated using classical Molecular Dynamics. These methodologies, usually termed QM/MM (Quantum Mechanics / Molecular Mechanics) interface [91], enable the analysis of some biological relevant events as the hydrolysis of peptide bond by proteases [92], hydrolysis of nucleoside triphosphates molecules in the active center of nucleosidase enzymes [93] or phosphorylation and dephosphorylation [94, 95, 96] processes, taking in account not only detailed aspects of formation and rupture of covalent bonds but also the movements of the entire protein associated to them [97].

Some of the effects of protein phosphorylation in protein structure must be studied taking into account that these effects are not stable but transient, allowing the presence of both phospho- and dephosphorylated states of the protein in a dynamic equilibrium [95]. In the near future, QM/MM interface techniques will not only the be able to simulate the behavior of the two extreme states of the protein, but also the intermediate situations of phosphate ions being added of removed from the surface of the macromolecules. Therefore, QM/MM interface will be able to analyze how these continuous processes can affect transient interactions with other proteins in signaling cascades [97] associated to immune response or diseases such as cancer.

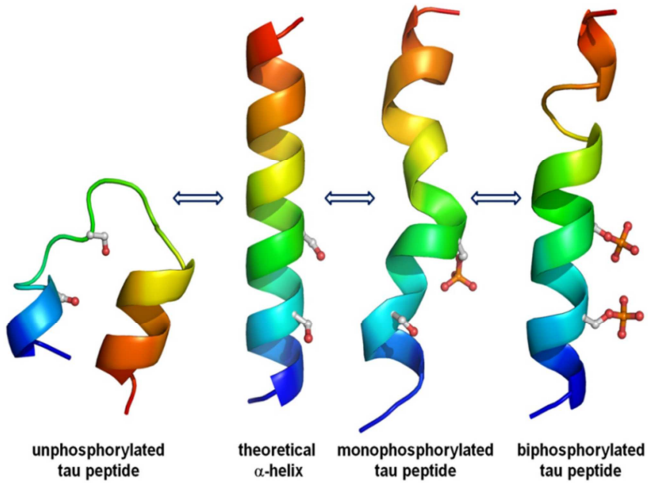


Figure 2. *Phosphorylation in protein structure and function.* Simulation, using Molecular Dynamics techniques, of the different equilibrium states of a peptide from the third tau microtubule-binding repeat. Phosphorylation displaces the pre-existing equilibrium towards a stabilized short alpha helical structure (modified from [84]).

2. Current proteomic strategies to study phosphorylated proteins and peptides

Phosphorylation plays important roles in regulating protein structure and function. Therefore, it also modulates protein catalytic activity, subcellular localization, and interaction processes [98, 99]. Protein reverse phosphorylation [100], concerts actions of kinases and phosphatases, and plays a central role in all processes, especially in cell signaling.

Phosphoproteomics involves the identification and quantification of phosphoproteins, and also the linkage of phosphosites to specific activation of the identified proteins. The analysis of the spatial and temporal aspects of protein phosphorylation is of great interest to the discovery of functions of specific biological processes.

2.1 *Phosphopeptides enrichment methods used previous to MS analysis*

For many approaches, including the study of immune disorders, the goal is to generate a global view of serine, threonine, and tyrosine phosphorylation within the sample, focusing specifically on the selected subset of phosphopeptides. Since the detection of phosphopeptides by MS is often hindered by suppression effects, many different strategies for removing unphosphorylated peptides have been established: immunoprecipitation by antibodies, pre-fraction systems such as ionic chromatographic exchange (SCX/SAX) and calcium phosphate precipitation, metal affinity chromatography e.g. IMAC, TiO_2, ZrO_2, and reverse phase chromatography (Figure 3).

Immunoprecipitation of phosphotyrosine is still much more frequent [101] than immunoprecipitation using phospho-serine or threonine antibodies. The reason is that affinity chromatography such as IMAC or titanium dioxide has higher capacity for phosphoserine and phosphothreonine peptide binding [102].

The most common techniques to enrich for individual and/or global phosphorylation are IMAC and Titanium Dioxide (TiO_2) [103], which are based on the high affinity of phosphate groups for metal ions such as Fe^{3+}, Zn^{+1}, Cu^{2+} and Ga^{3+}. The phosphopeptides that are negatively charged will bind the positively charged metal ions by electrostatic interactions. One of the main limitations associated with both phosphopeptide enrichments is the nonspecific retention of non-phosphorylated acidic peptides, due to the weak affinity between negatively charged carboxylate and positively charged metal ions. However, conversion of carboxylate groups to esters effectively eliminates nonspecific retention of non-phosphorylated peptides, although this has the drawback of increased complexity in the subsequent MS analysis. During the last five years, titanium dioxide (TiO_2) has emerged as the most common of the metal oxide affinity chromatography (MOAC) based phosphopeptide enrichment methods. This technique offers increased capacity compared to IMAC resins in order to bind and elute mono-phosphorylated peptides. TiO_2 exploits the same principle as IMAC, and is similarly prone to nonspecific retention of acidic non-phosphorylated peptides. However, when loading peptides in 2, 5-dihydroxybenzoic acid (DHB) [104], glycolic and phthalic acids, nonspecific binding to TiO_2 is reduced, thereby improving phosphopeptide enrichment without chemical modification of the sample. In addition, TiO_2 is often considered to be interchangeable with IMAC. It works on similar levels of sample amounts (e.g., micrograms of protein) for the identification of phospho-sites by MS analysis. Recently, SIMAC [105, 106] appeared as a phosphopeptide enrichment tool which is exploiting the properties of IMAC coupled to TiO_2, making it possible to carry out more refined studies.

Another phosphopeptide enrichment prior to mass spectrometric analysis is ZrO_2 [107] and its principle is based on metal affinity chromatography like IMAC and TiO_2. ZrO_2 allows isolation of single phosphorylated peptides more selectively than TiO_2 when using α and β casein as protein models. Indeed, it has been successfully used in the large-scale charac-

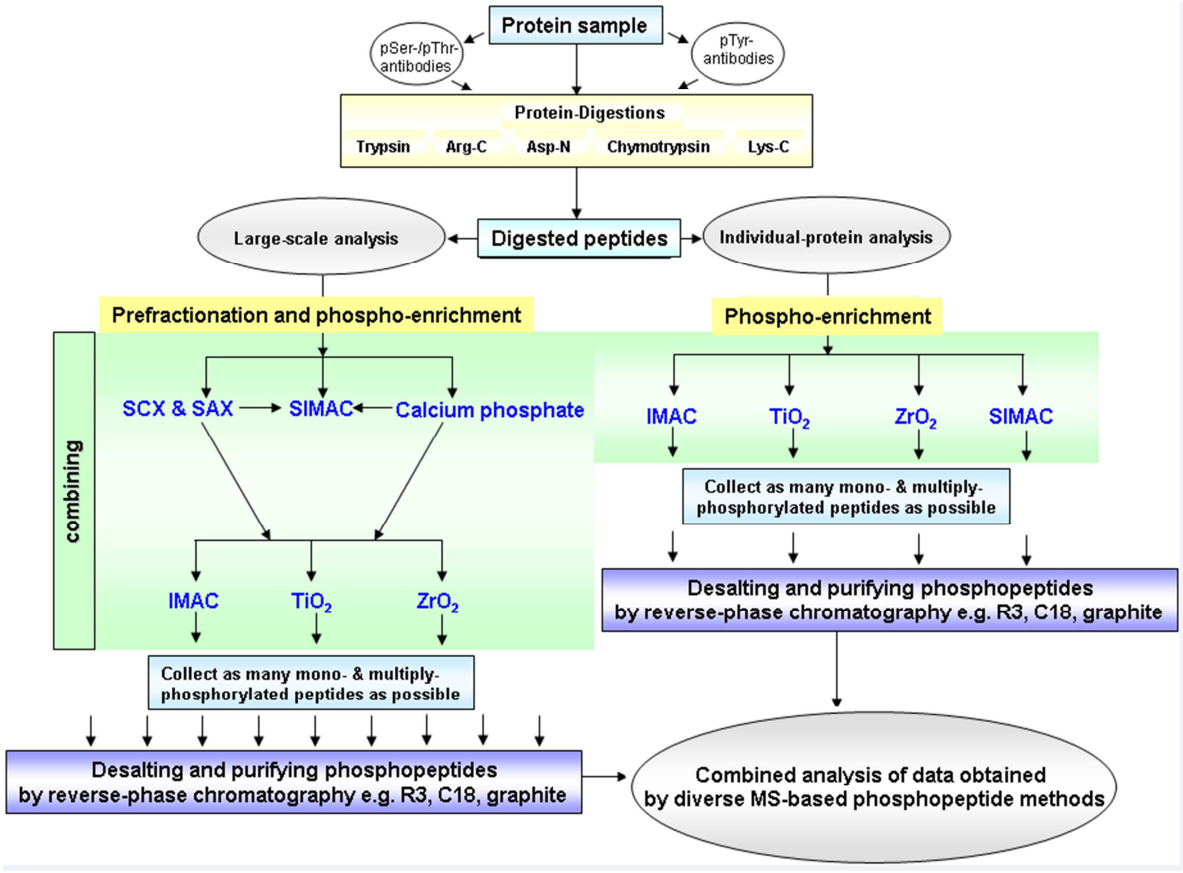


Figure 3. *Schematic workflow of current phosphoproteomic strategies.* Phosphorylated proteins obtained by immunoprecipitation from tissue or cell extract must be digested using one or a combination of different enzymes prior to bottom up MS.For a unique phosphoprotein analysis, IMAC, TiO2, ZrO2 or SIMAC are directly applied to the resulting peptides. For large scale phosphopeptide analysis, the peptides must be loaded onto pre-fraction columns (SCX/SAX, Calcium phosphate or SIMAC is used). In order to capture as many phosphopeptides as possible, it is necessary to use different phosphopeptide enrichments methods (IMAC, TiO2, ZrO2) when applying pre-fraction methods. Finally, the phosphopeptides which have been previously captured combining different phospho-strategies, must be purified, desalted and concentrated by reversed phase chromatography (POROs R3, C18 Disks and /or graphite) in order to be analyzed by MS.

terization of phosphoproteins.

Furthermore, a strategy which consists of fractionating and subsequently enriching phosphopeptides on a proteome wide scale is based on strong cation/anion exchange (SCX and SAX) chromatography. The principle of SCX/SAX phosphopeptide enrichment is based on the negative charged phosphate group (PO_4^-) of the phosphopeptides. In cation exchange chromatography, a positively charged analyte is attracted to a negatively charged solid-support, while in anion exchange chromatography negatively charged molecules are attracted to a positively charged solid-support. SAX has previously been successfully combined with IMAC [108] and resulted in greater recovery and identifications by MS of mono-phosphorylated peptides originating from membrane proteins. SCX has in a similar way been combined with IMAC (Fe^{3+}) and MS analysis, allowing the identification of thousands of phosphorylated residues from complex biological samples [9].

Calcium phosphate precipitation is a strategy that provides a useful pre-fractionation step to simplify and enrich phosphopeptides from complex samples. Zhang and co workers [109] have demonstrated that phosphopeptide precipitation by calcium phosphate combined with a two step IMAC (Fe^{3+}) procedure resulted in the observation of an increased number of phosphopeptides. This method consists of precipitating phosphopeptides by adding 0.5M $NaHPO_4$ and 2M NH_3OH to the peptide-mixture followed by 2M $CaCl_2$. The sample is vortexed and centrifuged, and, subsequently, the supernatant is removed before washing the pellet with 80 mM $CaCl_2$. The washed pellet is dissolved in 5% of formic acid and the resulting peptide mixture is desalted by reversed phase chromatography before isolating the phosphopeptides by IMAC (Fe^{3+}).

Finally, since most phosphopeptide analysis is nowadays performed by MS, and this technique is sensitive to contaminants such as salts, it is necessary to clean up the samples prior to analysis, generally by reversed phase chromatography combining POROs R3 with C18 Disks and also graphite powder [110, 111, 112]. Poros R3, C18 Disks and graphite powder are materials containing long hydrocarbon chains, proven to be effective for the desalting and cleaning of very hydrophilic peptides, including phosphopeptides [104, 113]. In 1999, Gobom and co-workers [110] introduced a micro column purification method in which a chromatographic resin was packed in the tip of a small constricted GELoader tip, creating a micro-column. Using a chromatographic approach, GELoader tips packed with R3, C18 or graphite material, contaminants like salts can be separated from the phosphopeptides. Indeed, using RP chromatography, molecules such as proteins, peptides and nucleic acids are separated according to their hydrophobicity. In addition to the removal of salts, these techniques also facilitate a concentration of the sample by the use of a low elution volume. This further improves the sensitivity and quality of the subsequent mass spectrometric analysis.

The phosphopeptide enrichments may be coupled with stable isotope labeling steps (such as SILAC, ICAT, iTRAQ and ^{18}O) for relative quantification and/or with AQUA analysis for absolute quantitative analysis [114, 115]. Since different metal ions appear to enrich for slightly different subsets of phosphorylated peptides, maximal coverage of the phosphoproteome may be obtained by multiple analyses with different metals, or by mixing multiple metal ions in a single enrichment step [26].

2.2 *MS analysis of phosphopeptides*

There have been important improvements in the MS technology concerning sampling throughput, sensitivity and mass accuracy over the past decade, but technical limitations in instrumentation still exist that make it impossible to fully sample the entire phosphoproteome. In addition, reproducibly and accurately measuring protein phosphorylation via MS presents many challenges [116]. Among these challenges is the fact that, many phosphorylated proteins are present in low overall abundance yet can have pronounced and important biological effects through their involvement in signaling cascades and other forms of downstream amplification such as transcriptional control. These properties make phosphoprotein or peptide enrichment a prerequisite for accurate analysis and increase sensitivity in MS detection. Robust sample lysis conditions to ensure stability of phosphoproteins that truly depict the phosphorylation state of proteins at the time of cell lysis are crucial. Furthermore, sample cleanup and pre-fractionation steps to crudely enrich for phosphopeptide pools and ensure buffer compatibility for phosphoenrichment protocols are also necessary [103, 117, 118].

Phosphorylation on serine and threonine residues are labile and conventional fragmentation CID (Collision Induced Dissociation) typically results in the partial neutral loss of phosphoric acid (H_3PO_4, 98/z) in MS^2 mode, due to the gas phase β-elimination of the phosphor-ester bond. Therefore, dehydroalanine and dehydroaminobutyric acid are generated. When peptide ions are fragmented by CID, series of y- and b-ions are formed [119, 120]. The peptide sequence is obtained by correlating mass difference between peaks in the y-ion series or between peaks in the b- ion series with amino acid residue masses. The CID fragmentation is occurring on the peptide backbone, and only limited sequence information is obtained. This event can also compromise the identification of phosphorylation sites. Related to phosphotyrosine residues, partial neutral loss is also observed (HPO_3, 80/z) in MS^2 mode, but the phosphate group on tyrosine residues is more stable than on serine and threonine residues. In addition, the phospho-finger-print characteristic of phosphotyrosine, is the phosphotyrosine immonium ion (~216 Da), which is a positive indicator for the presence of a peptide phosphorylated on tyrosine [121, 122].

The ion originating from neutral loss of phosphoric acid (H_3PO_4) can be selected for further fragmentation by MS^3 mode. The selected ion, after neutral loss fragmentation, is automatically selected for further fragmentation. This gives the possibility to add extra energy for the fragmentation of peptide backbone. Although, the MS^3 mode requires that the

selected ion is abundant in order to observe the fragmented ions. A pseudo-MS^3 development is MultiStage Activation (MSA) [123], which was implemented on quadrupole-IT and linear IT-orbitrap. In MSA, the fragmentation of the precursor ion occurs simultaneously with the fragmentation of the ion originating from the neutral loss. Then, the MS^2 and MS^3 mass-data are combined in a hybrid spectrum, resulting in improved sequence information and also in an improvement of the confidence for the phosphorylation site assignment.

Alternative fragmentations to CID are ECD (electron capture Dissociation) and ETD (Electron transfer dissociation). By ECD, radical peptide ions are obtained when multiply-charged peptide ions are rationing with low-energy thermal-electrons. In addition, this fragmentation occurs in the peptide between the backbone amide and the alpha carbon, generating c and z-ions [124]. An advantage of ECD is that it occurs only on the peptide backbone, and labile phosphate groups are left intact on the resulting c- and z- fragment ions, thus, enabling the identification of the specific phosphorylation sites. Therefore, it is extremely useful for the analysis of multiply-phosphorylated peptides. One disadvantage of ECD is that it has selectivity for disulfide bonds, due to the high radical affinity of the bond [125, 126]. The main drawback of ECD is only that it is used in the FT-ICR instruments due to the requirement of a static magnetic field for the thermal electrons, meaning high costs and high specialization.

By ETD, c- and z- ions are also generated. Actually, this fragmentation was developed in order to carry out ECD-like dissociation experiments, in a Quadrupole Linear Ion Trap [123, 127]. ETD is a chemical process in which reaction with fluoranthene radical anions disrupts the peptide backbone at regular intervals. ETD preserves the intact information about labile modifications, which are not observed directly when using CID. For instance, phosphate groups are good leaving groups, which mean that they are easily lost in the excitation process. However, by using ETD one can directly observe fragments that contain the intact phosphopeptides. The drawback of ETD is less sensitive compared to CID, because of lower ionization efficiency. As a result, we recommend using CID to start with, and would recommend switching to ETD in case you weren't able to determine the phosphorylation site.

2.3 *Database dependent search for the identification of phosphopeptides*

The new generation of mass spectrometers renders identification of thousands of proteins. However, proteins are normally identified based on several peptides, whereas phosphorylation sites are only identified by few peptides. Since different phosphorylation sites on the same protein usually behave differently [17], each phosphorylation site has to be evaluated individually. In addition, the identification of multi phosphorylated peptides has to be carefully evaluated because of the distinct behavior of individual phosphorylation sites. Identification of phosphopeptide sites involves database dependent search, application of filter parameters, estimation of false discovery rate and validation.

Care must be taken when defining the search parameters for a data set with spectra of phosphopeptides. The definition of phosphoserine and phosphothreonine modifications must include neutral loss of HPO_3 (79.966/z) and H_3PO_4 (97.977/z) whereas phosphotyrosine displays no characteristic neutral loss but instead a marker ion at 216/z. Furthermore, an apparent neutral loss of approximately 80/z and 98/z might have other causes than loss of HPO_3 or H_3PO_4 (Neutral Loss-Based Phosphopeptide Recognition: A Collection of Caveats, Journal of Proteome research, 2007, 6, 2866-2873). Moreover, the negative charge of a phospho group can shield some tryptic sites if located close to the cleavage site. The search parameters must therefore also allow for 2-3 missed tryptic cleavage sites.

The general available search engines are not optimized specifically for phosphopeptide identification and spectra of phosphopeptides are often of lower quality because of the labile phospho group. Therefore quality filtering using information from complementary spectra becomes essential. For example, complementary spectra of a specific phosphopeptide can be obtained by CID, HCD, ETD and MS^3. Furthermore, digesting the protein solution with alternative enzymes can provide alternative peptides that may ionize or fragment better. Ideally the software used is able to combine information from all the above mention spectra sources and provide a filtered list with high quality results.

The decoy approach can be used to estimate false discovery rate. The decoy approach consists of searching two sequence databases which contains the protein sequences in both forward and reversed (decoy) orientations, respectively. In addition, the number of false-positive hits that are found in the decoy database, is then used to predict the number of false positives found in the forward database. The decoy approach has been described as a useful method for analyzing LC-MS/MS data from yeast [128, 129], and later, it has also been used for phosphoproteome analysis [130, 131]. Search engines like Mascot [132] and VEMS [133] have built in false discovery rate on the peptide level which can be used for filtering. However, no search engine is able to provide a false discovery rate based on combined data from the different spectra sources mentioned above.

Finally, the phosphopeptide assignments must be validated. Validation can for example be done by obtaining spectra on a synthetic phosphopeptide standard which has the same phosphorylation sites as a phosphopeptides of interest from the biological sample.

It can, in some cases, be difficult to assign the correct phoshporylation sites in a peptide containing many serines, threonines and tyrosines. To this end, VEMS has a spectrum viewer that can generate all possible phophorylation sites for a peptide and overlay the theoretical fragment ions from these with the observed fragment ions [134]. The possible phosphorylation sites are restricted by the observed parent ion mass and the identified peptide sequence. This tool is

useful for manually validating that the correct phosphorylation sites in the peptides are annotated. It also allows detection of spectra originating from two phosphopeptides with the same amino acid sequence but different phosphorylation sites that could not be separated on the reverse phase column. VEMS considers all fragment ion types characteristic for phosphopeptides including multiple combinations of losses of -79.966/z and -97.977/z for the scoring.

The software MSQuant (http://msquant.sourceforge.net) [17, 18] is compatible with MS^2 and MS^3 searches performed using Mascot. The MSQuant scoring is probability based and takes into account the number of matched b- and y-ions for all possible combinations of phosphorylation sites in a given peptide sequence. Another program, Ascore (http://ascore.med.harvard.edu/ascore.php), uses a probability-based scoring algorithm to validate phosphopeptides assigned by SEQUEST [130, 131].

A recent comparison of the database dependent search engines Mascot [132], SEQUEST [135], OMSSA [136], and X!Tandem [137] suggested that it is useful to combine results from several search engines [138]. The comparison indicated that, when searching high mass accuracy data sets, Mascot and SEQUEST obtained higher sensitivity for phosphopeptide identifications [138]. This study also demonstrated that replicate-runs of the same sample significantly increased the total number of phosphopeptide identifications. Although, a particular search algorithm may perform better under certain conditions, researchers should not limit their searches to using just one algorithm since a combined use of search engines offer better coverage [139]. Recent alternative strategies for database dependent search are based on peptide sequence tags e.g. GutenTag [140] and InsPecT [141]. However, the sequence tag based methods have so far not been applied to large scale phosphopeptide analysis. The sequence tag based methods can become important in the future when the quality of phosphopeptide spectra improves.

3. MS-based quantitative strategies for phosphopeptides

Proteins containing amino acids with one or more of the stable isotopes of ^2H, ^{13}C, ^{15}N or ^{18}O can be used as internal standards by addition, at an early stage of the analysis, of a complex protein sample. There are two approaches for introducing a stable isotope into a proteins or peptides: metabolic labeling using whole cells grown in culture (e.g. SILAC) or chemical labeling (e.g. iTRAQ, ICAT).

Measuring the changes in phosphorylation is critical for understanding the biology of a phosphorylation event, since protein phosphorylation is very dynamic and constantly changing throughout the life of a cell. We restrict the discussion here to three MS-based quantitation strategies which have direct utility towards measuring changes in protein phosphorylation: SILAC, iTRAQ, and AQUA. Other chemical labeling techniques that rely on stable isotope incorporation using e.g. ^{18}O labeled water during trypsin digestions and stable isotope incorporation ICAT can also be considered with relevant information, but will not be described here.

3.1 *SILAC*

Stable isotope labeling by amino acids in cell culture (SILAC) is a quantitative method based on in vivo labeling of proteins in cell cultures with amino acids that contain stable isotopes (non radioactive, e.g. ^2H, ^{13}C and ^{15}N) [115, 142]. In its simplest form, two separated cell cultures are grown in a pair-wise fashion; for example, culture A might be yeast cells grown under "normal" conditions (light conditions) while culture B might be yeast cells grown in the presence of a stress condition. The growth conditions of the cells are identical (except for the presence of the stress-stimuli), but the growth media of culture B has an essential amino acid (one not synthesized by the cell) replaced with an isotopically "heavy" form of that amino acid (e.g. $^{13}C_6$-arginine). A number of cell lines have been used for SILAC experiments, and the growth and morphology of the cells have not been affected by the isotopically labeled amino acid [9, 142, 143]. After approximately five rounds of doubling, cellular proteins are essentially 100% labeled with the selected amino acid. After culturing, the light and heavy cell populations are combined (1:1) into one pool and the proteins are isolated. The protein pool is then digested with a protease, typically trypsin, to form a peptide pool that is analyzed by MS. Each peptide analyzed will be present in two forms, the light and the heavy form. The two forms have the same chemical properties, so they have approximately the same chromatographic retentions, ionization efficiencies, and fragmentation characteristics, but they are distinguishable based on the mass difference due to the heavy isotope incorporation in the selected amino acid. The peak signals produced by the light and heavy forms of a peptide are measured by the mass spectrometer, and a relative quantification of that peptide from the two cultures is calculated. Tandem MS is also performed in the same experiment on either the light or the heavy form, therefore, the identity of the peptide and the protein is determined. Indeed, all peptides, both phosphorylated and non-phosphorylated, that contain the isotopically labeled amino acid, are available for relative quantification by SILAC. The SILAC method is compatible with the above mentioned enrichment of phosphoproteins/phosphopeptides including the immunoprecipitation of a target protein [144]. To assist with the enrichment of phosphopeptides in the SILAC method, the combination of SCX chromatography and IMAC, have been employed after proteolytic digestion [9]. This approach enriches the phosphopeptides and helps removing non-phosphorylated peptides that can act as noise in the quantification experiment (Figure 4A).

3.2 *iTRAQ*

The second method for the global quantification of proteins and protein modifications is an in vitro chemical labeling procedure called iTRAQ. The iTRAQ reagent consists of two to eight isobaric (same nominal mass) tags that can be used to label two to eight separate protein samples; for example, one sample might be "normal" yeast cells while the three remaining samples might be yeast cells grown at

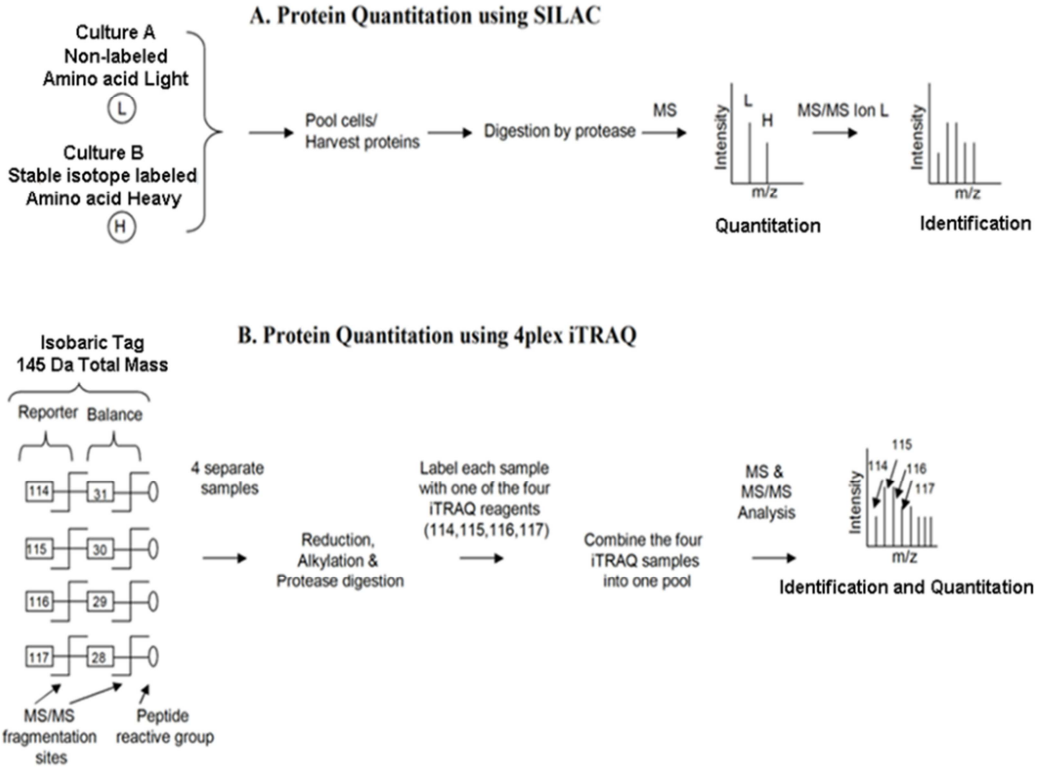


Figure 4. Scheme of two quantitative proteomic methods based on isotopic labeling. *(A) Protein quantification by SILAC.* Stable isotope labeling with amino acids in cell culture is an approach for the in vivo incorporation of a tag into proteins for relative quantitation by MS. The method relies on the incorporation of amino acids with substituted stable isotopic nuclei (e.g. 2H, 13C, 15N). Two cell populations are grown in culture media that are identical except that one of them contains a "light, L" and the other a "heavy, H" form of a particular amino acid (e.g. 12C and 13C labeled L-lysine, respectively). After a number of cell divisions, each instance of this particular amino acid will be replaced by its isotope labeled analog, and the quantitation information is provided in the MS mode by the intensity peaks observed in the spectra. *(B) Protein quantification scheme of the iTRAQ method.* This method is relying on the derivatization of primary amino groups in intact proteins using isobaric tag for relative and absolute quantitation. Due to the isobaric mass design of the iTRAQ reagents, differentially labeled proteins do not differ in mass; accordingly, their corresponding proteolytic peptides appear as single peaks in MS scans. Because quantitative information is provided by isotope-encoded reporter ions that can only be observed in MS/MS spectra. In addition phosphorylation analysis by MS and MS/MS can be carried out on iTRAQ labeled peptides to obtain quantitative phosphopeptide information.

three different concentrations of a stress treatment. The iTRAQ tags contain three regions: a peptide reactive region, a reporter region, and a balance region [145]. The peptide reactive region of the tag consists of an NHS ester and is designed to react with the N-termini and lysines of peptides after protease digestions. In the case of 4-plex iTRAQ, the four reporter groups appear in the tandem mass spectrum at m/z 114, 115, 116, and 117. The attached balance groups are designed to make the total mass of the balance and reporter group 145 Da for each tag, resulting in balance groups of 31 Da, 30 Da, 29 Da, and 28 Da, respectively. Protein samples for quantification are separately isolated and digested proteolytically, and each sample is chemically labeled with one of the iTRAQ reagents. After labeling, the samples are combined and subsequently analyzed by MS. As the iTRAQ reagents are isobaric, identical peptides from each sample will have identical masses, therefore there is no division of the precursor signals in the first stage of mass analysis that could lead to increased spectral complexity by the combination of multiple samples. Additionally, the isobaric nature of the reagent increases the ion population for a given peptide by summarizing the amount of a peptide from each sample. This makes the detection of the peptides more sensitive. During tandem MS, fragmentation takes place along the peptide backbone and also between the reporter and balance region of the tag which facilitates the quantitation based on the intensity of the reporter ions. The reporter ions in the tandem mass spectrum are in general more intense than the fragment ions. The relative amounts of these reporter ions correspond to the relative amounts of the peptides present in the four samples. In contrast to SILAC and AQUA (described below), it is during tandem MS experiments, and not during the first stage of mass analysis, that relative quantification of peptides takes place. Phosphoproteins can be analyzed in an identical manner as well as non-phosphorylated proteins with the iTRAQ methodology. The iTRAQ reagent labels phosphopeptides to the same degree as non-phosphorylated peptides and it does not affect the stability of phosphopeptides. Enrichment strategies, such as IMAC [146, 147] or immunoprecipitation with anti-phosphotyrosine antibodies [147], have been utilized to remove non-phosphorylated peptides to focus the analysis on site-specific phosphorylation. Also, since iTRAQ is an in vitro labeling procedure it can be applied to clinical samples such as tumor tissues and fluids (e.g.

serum, urine, blood). iTRAQ has been described as a very powerful method for the quantification of phosphorylation on a proteomic scale. In addition, White and co-workers [148] applied iTRAQ combined with MRM for phospho quantitative analysis of signaling networks, identifying and quantifying 222 tyrosine phosphorylated peptides, getting an extremely high percentage of signaling nodes covered (Figure 4B).

3.3 *AQUA*

The AQUA strategy provides an absolute quantification of a protein of interest [149] In the AQUA method, a peptide from the protein of interest is constructed synthetically containing stable isotopes, and the isotopically labeled synthetic peptide is called AQUA peptide. The synthetic peptides can be synthesized with modifications such as phosphorylation to allow for the direct, quantitative analysis of post-translationally modified proteins. The stable isotopes are incorporated into the AQUA peptide by using isotopically "heavy" amino acids during the synthesis process of the interesting peptide (native peptide). Therefore, the synthetic peptide has a mass increase of e.g. 10 Daltons, due to the incorporation of a $^{13}C_6$ and $^{15}N_4$-arginine into the synthetic peptide, compared to the native peptide. Although the mass difference between the native and the synthetic peptide allows the mass spectrometer to differentiate between the two forms, both forms have the same chemical properties, resulting in the same chromatographic retention, ionization efficiency, and fragmentation distribution. In AQUA experiments, a known amount of the isotopically labeled peptide is added to a protein mixture, which is proteolytically digested, and later analyzed by MS. Since the native peptide and its synthetic counterpart have the same chemical properties, the MS signal from the quantified synthetic peptide can be compared to the signal of the native peptide. This finally allows for the absolute quantification of the peptide to be determined [150]. Multiple AQUA peptides can be used to quantify multiple proteins in a single experiment (Figure 5).

4. Concluding remarks and future perspectives

Cancer and immune disorders, remain among the leading causes of death worldwide. Therefore, the identification of useful biomarkers and the improvement of the understanding of the development of these diseases is still a need.

The immune system is easily influenced by the existence of cancer in the body, even when it is in a preclinical stage, and these studies should be expanded and extended in the future to answer the numerous questions concerning (a) the roles of immune cells in cancer surveillance (b) the characteristics of inflammation associated with cancer development, (c) the effects of environment/lifestyle factors on the immune system, and (d) the interaction between aging and immunity in the occurrence of cancer and other diseases.

The importance of protein kinase-regulated signal transduction pathways in immunology disorders and cancer has led to the development of drugs that inhibit protein kinases at the apex or intermediary levels of these pathways. Protein phosphorylation assignment studies of these signaling path-

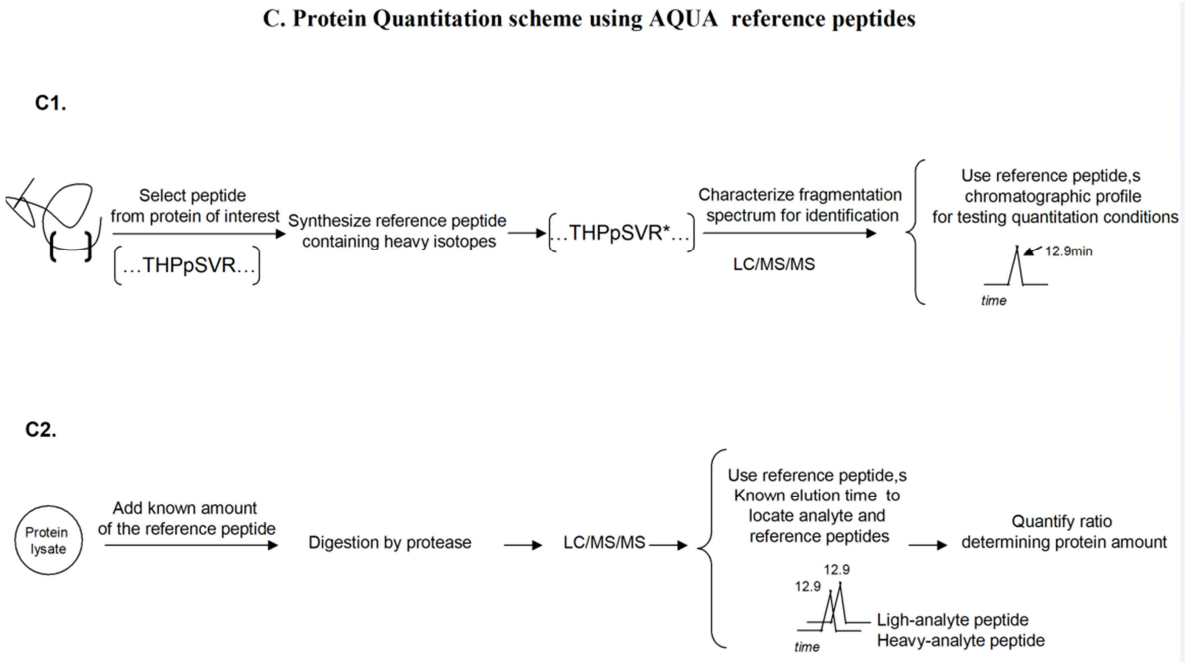


Figure 5. Scheme of absolute quantitative proteomic method based on the synthesis of reference peptides with stable isotopes. *(C1)* The first steps of the AQUA method consist of synthesizing peptides of interest with incorporated stable isotopes as ideal internal standards to mimic native peptides formed by proteolysis. These synthetic peptides can also be prepared with covalent modifications, like phosphorylation, that are chemically identical to naturally occurring posttranslational modifications. *(C2)* AQUA internal standard peptides are added to the protein sample, after proteolysis, and used to precisely and quantitatively measure the absolute levels of proteins and their post-translational modifications by MS.

ways will provide important insights into the operation and connectivity of these pathways that will facilitate the identification of the best targets for cancer therapies (e.g. the identification of a phosphate group on a specific serine, threonine or tyrosine by phospho enrichments combined with MS). Moreover, phosphoproteomic analysis of individual tumors will help match targeted cancer drugs to the appropriate patients.

The current phosphoproteomic goals imply the identification of phosphoproteins, mapping of phosphorylation sites, quantitation of phosphorylation under different conditions, and the determination of the stoichiometry of the phosphorylation. In addition, knowing when a protein is phosphorylated, which kinase/s is-are involved, and how each phosphorylation fits into the signaling network, are also important challenges for researchers in order to understand the significance of different biological events.

The new MS technologies are fundamental for cataloging all this information, and it is heading towards collecting accurate data on phosphopeptides on a global scale. Finally, it must be pointed out the possible difficulties to get sufficient amount of specific phosphorylated proteins of specific low abundant protein-kinases in vivo which might limit the usability of the phosphoproteome analysis.

Acknowledgements

Support for EL provided by MICINN (Ministerio de Ciencia e Innovación de España) for Phosphoproteomics. Support for IL provided by Grant "Leukemia & Linfoma Funding Marcos Fernández (Fermoselle)" for Hematology Research at the Hosp. Univ. 12 de Octubre 2010-2011. Support for RM was provided by Fundação para a Ciência e a Tecnologia grant "PTDC/QUI-BIQ/099457/2008". PG-P is supported by grants SAF2007-61926 from the Spanish Ministry of Education and Science; by CIBER Fisiopatología de la Obesidad y Nutrición, an initiative of the Instituto de Salud Carlos III ; by grant S-BIO-0260/2006-COMBACT from the Comunidad de Madrid; and by grant FP7-223431 (EU project "Divinocell"). Financial support from the "Fundación Ramón Areces" to CBMSO is also acknowledged. JM and JJW were recipients of Torres-Quevedo post-doctoral fellowships financed by the European Social Fund and the Ministerio de Educación y Ciencia. Work at Biomol-Informatics S.L. is partially financed by the European Social Fund.

References

1. G. Burnett, E.P. Kennedy, J. Biol. Chem. 211 (1954) 969-980.
2. W. Eckhart, M.A. Hutchinson, T. Hunter, Cell. 18 (1979) 925-933.
3. S. Shoji, D.C. Parmelee, R.D. Wade, S. Kumar, L.H. Ericsson, K.A. Walsh, H. Neurath, G.L. Long, J.G. Demaille, E.H. Fischer, K. Titani, Proc. Natl. Acad. Sci. U. S. A. 78 (1981) 848-851.
4. L.B. Ray, T.W. Sturgill, Proc. Natl. Acad. Sci. U. S. A. 84 (1987) 1502-1506.
5. N.K. Tonks, H. Charbonneau, C.D. Diltz, E.H. Fischer, K.A. Walsh, Biochemistry (Mosc.). 27 (1988) 8695-8701.
6. P.L. Schwartzberg, S.P. Goff, E.J. Robertson, Science. 246 (1989) 799-803.
7. K. Shah, Y. Liu, C. Deirmengian, K.M. Shokat, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 3565-3570.
8. S.B. Ficarro, M.L. McCleland, P.T. Stukenberg, D.J. Burke, M.M. Ross, J. Shabanowitz, D.F. Hunt, F.M. White, Nat. Biotechnol. 20 (2002) 301-305.
9. A. Gruhler, J.V. Olsen, S. Mohammed, P. Mortensen, N.J. Faergeman, M. Mann, O.N. Jensen, Mol. Cell. Proteomics. 4 (2005) 310-327.
10. G. Manning, D.B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, Science. 298 (2002) 1912-1934.
11. H. Daub, J.V. Olsen, M. Bairlein, F. Gnad, F.S. Oppermann, R. Korner, Z. Greff, G. Keri, O. Stemmann, M. Mann, Mol. Cell. 31 (2008) 438-448.
12. N. Dephoure, C. Zhou, J. Villen, S.A. Beausoleil, C.E. Bakalarski, S.J. Elledge, S.P. Gygi, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 10762-10767.
13. C.R. Ingrell, M.L. Miller, O.N. Jensen, N. Blom, Bioinformatics. 23 (2007) 895-897.
14. O.N. Jensen, Nat. Rev. Mol. Cell Biol. 7 (2006) 391-403.
15. C. Kumar, M. Mann, FEBS Lett. 583 (2009) 1703-1712.
16. R. Matthiesen, O.N. Jensen, Methods Mol. Biol. 453 (2008) 105-122.
17. J.V. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen, M. Mann, Cell. 127 (2006) 635-648.
18. J.V. Olsen, M. Mann, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 13417-13422.
19. P.P. Ahern, A. Izcue, K.J. Maloy, F. Powrie, Immunol. Rev. 226 (2008) 147-159.
20. L. Romani, Med. Mycol. 46 (2008) 515-529.
21. S. Schneider-Schaulies, J. Schneider-Schaulies, Curr. Top. Microbiol. Immunol. 330 (2009) 243-269.
22. E. Codner, Pediatr. Endocrinol. Rev. 6 (2008) 228-234.
23. K.A. Kim, M.S. Lee, Front. Biosci. 14 (2009) 657-664.
24. A. Prawan, B. Buranrat, U. Kukongviriyapan, B. Sripa, V. Kukongviriyapan, J. Cancer Res. Clin. Oncol. 135 (2009) 515-522.
25. P. Cohen, Curr. Opin. Cell Biol. 21 (2009) 317-324.
26. K. Ashman, E.L. Villar, Clin. Transl. Oncol. 11 (2009) 356-362.
27. I. Dolado, A.R. Nebreda, Cancer Cell. 14 (2008) 427-429.
28. M. Malumbres, M. Barbacid, Nat. Rev. Cancer. 1 (2001) 222-231.
29. A. Swat, I. Dolado, J.M. Rojas, A.R. Nebreda, Mol. Cell. Biol. 29 (2009) 3332-3343.
30. J.J. Ventura, S. Tenbaum, E. Perdiguero, M. Huth, C. Guerra, M. Barbacid, M. Pasparakis, A.R. Nebreda, Nat. Genet. 39 (2007) 750-758.
31. J. Hulit, K. Suyama, S. Chung, R. Keren, G. Agiostratidou, W. Shan, X. Dong, T.M. Williams, M.P. Lisanti, K. Knudsen, R.B. Hazan, Cancer Res. 67 (2007) 3106-3116.
32. M.D. Lu, Y. Wang, L. Chen, J. Qin, P. Li, X.P. Cui, A.G. Shen, Zhonghua Bing Li Xue Za Zhi. 36 (2007) 840-841.
33. L. Moro, A.A. Arbini, E. Marra, M. Greco, Int. J. Oncol. 30 (2007) 217-224.
34. A. Zebisch, A.P. Czernilofsky, G. Keri, J. Smigelskaite, H. Sill, J. Troppmair, Curr. Med. Chem. 14 (2007) 601-623.
35. K. Matsuura, Y. Nohno, N. Hijiya, T. Uchida, Y. Tsukamoto, M. Moriyama, Pathol. Int. 56 (2006) 368-374.
36. V. Gerke, S.E. Moss, Biochim. Biophys. Acta. 1357 (1997) 129-154.
37. L.C. Alldridge, H.J. Harris, R. Plevin, R. Hannon, C.E. Bryant, J. Biol. Chem. 274 (1999) 37620-37628.
38. T. Sudo, H. Hidaka, J. Biol. Chem. 273 (1998) 6351-6357.

39. R. Araki, R. Fukumura, A. Fujimori, Y. Taya, Y. Shiloh, A. Kurimasa, S. Burma, G.C. Li, D.J. Chen, K. Sato, Y. Hoki, K. Tatsumi, M. Abe, Cancer Res. 59 (1999) 3543-3546.
40. R. Boutros, V. Lobjois, B. Ducommun, Nat. Rev. Cancer. 7 (2007) 495-507.
41. K. Galaktionov, A.K. Lee, J. Eckstein, G. Draetta, J. Meckler, M. Loda, D. Beach, Science. 269 (1995) 1575-1577.
42. R.Q. Chen, Q.K. Yang, B.W. Lu, W. Yi, G. Cantin, Y.L. Chen, C. Fearns, J.R. Yates, 3rd, J.D. Lee, Cancer Res. 69 (2009) 2663-2668.
43. S. Tsukada, D.C. Saffran, D.J. Rawlings, O. Parolini, R.C. Allen, I. Klisak, R.S. Sparkes, H. Kubagawa, T. Mohandas, S. Quan, J.W. Belmont, M.D. Cooper, M.E. Conley, O.N. Witte, Cell. 72 (1993) 279-290.
44. D. Vetrie, I. Vorechovsky, P. Sideras, J. Holland, A. Davies, F. Flinter, L. Hammarstrom, C. Kinnon, R. Levinsky, M. Bobrow, C.I.E. Smith, D.R. Bentley, Nature. 361 (1993) 226-233.
45. T. Futatani, T. Miyawaki, S. Tsukada, S. Hashimoto, T. Kunikata, S. Arai, M. Kurimoto, Y. Niida, H. Matsuoka, Y. Sakiyama, T. Iwata, S. Tsuchiya, O. Tatsuzawa, K. Yoshizaki, T. Kishimoto, Blood. 91 (1998) 595-602.
46. C.I. Smith, B. Baskin, P. Humire-Greiff, J.N. Zhou, P.G. Olsson, H.S. Maniar, P. Kjellen, J.D. Lambris, B. Christensson, L. Hammarstrom, D. Bentley, D. Vetrie, K.B. Islam, I. Vorechovsky, P. Sideras, J. Immunol. 152 (1994) 557-565.
47. C.I.E. Smith, A.B. Satterthwaite, O.N. Witte, in: H.D. Ochs, C.I.E. Smith (Eds.), Primary Immunodeficiency Diseases a Molecular and Genetic Approach, Oxford University Press, New York, 2007 pp. 279-303.
48. A.J. Mohamed, L. Yu, C.M. Backesjo, L. Vargas, R. Faryal, A. Aints, B. Christensson, A. Berglof, M. Vihinen, B.F. Nore, C.I. Smith, Immunol. Rev. 228 (2009) 58-73.
49. S. Thieffry, M. Arthuis, J. Aicardi, G. Lyon, Rev. Neurol. (Paris). 105 (1961) 390-405.
50. C.D. West, R. Hong, N.H. Holland, J. Clin. Invest. 41 (1962) 2054-2064.
51. L. Yel, J. Clin. Immunol. 30 (2010) 10-16.
52. P. Brandtzaeg, G. Karlsson, G. Hansson, B. Petruson, J. Bjorkander, L.A. Hanson, Clin. Exp. Immunol. 67 (1987) 626-636.
53. T. Klemola, J. Pediatr. Gastroenterol. Nutr. 7 (1988) 537-543.
54. C. Cunningham-Rundles, Adv. Hum. Genet. 19 (1990) 235-266.
55. A. Ferreira, M.C. Garcia Rodriguez, M. Lopez-Trascasa, D. Pascual Salcedo, G. Fontan, Clin. Immunol. Immunopathol. 47 (1988) 199-207.
56. W. Strober, M.C. Sneller, Ann. Allergy. 66 (1991) 363-375.
57. C. Cunningham-Rundles, in: E.R. Stiehm, H.D. Ochs, J.A. Winkelstein (Eds.), Immunologic Disorders in Infanst & Children, Elsevier Saunders, Philadelphia, 2004 pp. 427-446.
58. E.G. De la Concha, M. Fernandez-Arquero, L. Gual, P. Vigil, A. Martinez, E. Urcelay, A. Ferreira, M.C. Garcia-Rodriguez, G. Fontan, J. Immunol. 169 (2002) 4637-4643.
59. L. Gual, A. Martinez, M. Fernandez-Arquero, M.C. Garcia-Rodriguez, A. Ferreira, G. Fontan, E.G. de la Concha, E. Urcelay, Tissue Antigens. 64 (2004) 671-677.
60. A. Martinez, L. Gual, M. Fernandez-Arquero, A. Nogales, A. Ferreira, M.C. Garcia-Rodriguez, G. Fontan, E.G. de la Concha, Genes Immun. 4 (2003) 316-320.
61. J. Mohammadi, R. Ramanujam, S. Jarefors, N. Rezaei, A. Aghamohammadi, P.K. Gregersen, L. Hammarstrom, J. Clin. Immunol. 30 (2010) 138-143.
62. O. Olerup, C.I. Smith, J. Bjorkander, L. Hammarstrom, Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 10653-10657.
63. N.D. Pozo, L.M. Medrano, M.C. Cenit, M. Fernandez-Arquero, A. Ferreira, M.C. Garcia-Rodriguez, E.G. de la Concha, E. Urcelay, C. Nunez, Hum. Immunol. 71 (2010) 861-864.
64. I. Vorechovsky, M. Cullen, M. Carrington, L. Hammarstrom, A.D. Webster, J. Immunol. 164 (2000) 4408-4416.
65. F.M. Schaffer, R.C. Monteiro, J.E. Volanakis, M.D. Cooper, Immunodefic. Rev. 3 (1991) 15-44.
66. F.M. Schaffer, J. Palermos, Z.B. Zhu, B.O. Barger, M.D. Cooper, J.E. Volanakis, Proc. Natl. Acad. Sci. U. S. A. 86 (1989) 8015-8019.
67. E. Castigli, S.A. Wilson, L. Garibyan, R. Rachid, F. Bonilla, L. Schneider, R.S. Geha, Nat. Genet. 37 (2005) 829-834.
68. R. Lopez-Mejias, N. del Pozo, M. Fernandez-Arquero, A. Ferreira, M.C. Garcia-Rodriguez, E.G. de la Concha, G. Fontan, E. Urcelay, A. Martinez, C. Nunez, Tissue Antigens. 74 (2009) 42-45.
69. Q. Pan-Hammarstrom, U. Salzer, L. Du, J. Bjorkander, C. Cunningham-Rundles, D.L. Nelson, C. Bacchelli, H.B. Gaspar, S. Offer, T.W. Behrens, B. Grimbacher, L. Hammarstrom, Nat. Genet. 39 (2007) 429-430.
70. E.G. De la Concha, M. Fernandez-Arquero, P. Vigil, F. Lazaro, A. Ferreira, M.C. Garcia-Rodriguez, G. Fontan, Tissue Antigens. 55 (2000) 359-363.
71. N. del Pozo, R. Lopez-Mejias, M. Fernandez-Arquero, A. Ferreira, M.C. Garcia-Rodriguez, E.G. de la Concha, G. Fontan, E. Urcelay, C. Nunez, Hum. Immunol. 70 (2009) 950-952.
72. K. Haimila, E. Einarsdottir, A. de Kauwe, L.L. Koskinen, Q. Pan-Hammarstrom, T. Kaartinen, K. Kurppa, F. Ziberna, T. Not, S. Vatta, A. Ventura, I.R. Korponay-Szabo, R. Adany, Z. Pocsai, G. Szeles, E. Dukes, K. Kaukinen, M. Maki, S. Koskinen, J. Partanen, L. Hammarstrom, P. Saavalainen, Genes Immun. 10 (2009) 151-161.
73. R. Lopez-Mejias, A. Martinez, N. Del Pozo, M. Fernandez-Arquero, A. Ferreira, E. Urcelay, G. Fontan, E.G. de la Concha, C. Nunez, Hum. Immunol. 69 (2008) 301-305.
74. J. Ortiz, M. Fernandez-Arquero, E. Urcelay, R. Lopez-Mejias, A. Ferreira, G. Fontan, E.G. de la Concha, A. Martinez, BMC Med Genet. 7 (2006) 56.
75. R.C. Ferreira, Q. Pan-Hammarstrom, R.R. Graham, V. Gateva, G. Fontan, A.T. Lee, W. Ortmann, E. Urcelay, M. Fernandez-Arquero, C. Nunez, G. Jorgensen, B.R. Ludviksson, S. Koskinen, K. Haimila, H.F. Clark, L. Klareskog, P.K. Gregersen, T.W. Behrens, L. Hammarstrom, Nat. Genet. 42 (2010) 777-780.
76. C. Picard, A. Puel, M. Bonnet, C.L. Ku, J. Bustamante, K. Yang, C. Soudais, S. Dupuis, J. Feinberg, C. Fieschi, C. Elbim, R. Hitchcock, D. Lammas, G. Davies, A. Al-Ghonaium, H. Al-Rayes, S. Al-Jumaah, S. Al-Hajjar, I.Z. Al-Mohsen, H.H. Frayha, R. Rucker, T.R. Hawn, A. Aderem, H. Tufenkeji, S. Haraguchi, N.K. Day, R.A. Good, M.A. Gougerot-Pocidalo, A. Ozinsky, J.L. Casanova, Science. 299 (2003) 2076-2079.
77. S.M. Holland, F.R. DeLeo, H.Z. Elloumi, A.P. Hsu, G. Uzel, N. Brodsky, A.F. Freeman, A. Demidowich, J. Davis, M.L. Turner, V.L. Anderson, D.N. Darnell, P.A. Welch, D.B. Kuhns, D.M. Frucht, H.L. Malech, J.I. Gallin, S.D. Kobayashi, A.R. Whitney, J.M. Voyich, J.M. Musser, C. Woellner, A.A. Schaffer, J.M. Puck, B. Grimbacher, N. Engl. J. Med. 357 (2007) 1608-1619.
78. Y. Minegishi, M. Saito, S. Tsuchiya, I. Tsuge, H. Takada, T. Hara, N. Kawamura, T. Ariga, S. Pasic, O. Stojkovic, A. Metin, H. Karasuyama, Nature. 448 (2007) 1058-1062.
79. Y. Minegishi, M. Saito, T. Morio, K. Watanabe, K. Agematsu, S. Tsuchiya, H. Takada, T. Hara, N. Kawamura, T. Ariga, H. Kaneko, N. Kondo, I. Tsuge, A. Yachie, Y. Sakiyama, T. Iwata, F. Bessho, T. Ohishi, K. Joh, K. Imai, K. Kogawa, M. Shinohara, M. Fujieda, H. Wakiguchi, S. Pasic, M. Abinun, H.D. Ochs, E.D. Renner, A. Jansson, B.H. Belohradsky, A. Metin, N.

Shimizu, S. Mizutani, T. Miyawaki, S. Nonoyama, H. Karasuyama, Immunity. 25 (2006) 745-755.
80. K.R. Engelhardt, S. McGhee, S. Winkler, A. Sassi, C. Woellner, G. Lopez-Herrera, A. Chen, H.S. Kim, M.G. Lloret, I. Schulze, S. Ehl, J. Thiel, D. Pfeifer, H. Veelken, T. Niehues, K. Siepermann, S. Weinspach, I. Reisli, S. Keles, F. Genel, N. Kutukculer, Y. Camcioglu, A. Somer, E. Karakoc-Aydiner, I. Barlan, A. Gennery, A. Metin, A. Degerliyurt, M.C. Pietrogrande, M. Yeganeh, Z. Baz, S. Al-Tamemi, C. Klein, J.M. Puck, S.M. Holland, E.R. McCabe, B. Grimbacher, T.A. Chatila, J. Allergy Clin. Immunol. 124 (2009) 1289-1302.
81. L.D. Notarangelo, A. Fischer, R.S. Geha, J.L. Casanova, H. Chapel, M.E. Conley, C. Cunningham-Rundles, A. Etzioni, L. Hammartrom, S. Nonoyama, H.D. Ochs, J. Puck, C. Roifman, R. Seger, J. Wedgwood, J. Allergy Clin. Immunol. 124 (2009) 1161-1178.
82. X. Zhao, I.R. Leon, S. Bak, M. Mogensen, K. Wrzesinski, K. Hojlund, O.N. Jensen, Mol. Cell. Proteomics. (2010) in press.
83. A.C. Roque, C.R. Lowe, Biotechnol. Bioeng. 91 (2005) 546-555.
84. M.B. Yaffe, S.J. Smerdon, Annu. Rev. Biophys. Biomol. Struct. 33 (2004) 225-244.
85. D. Chevalier, E.R. Morris, J.C. Walker, Annu. Rev. Plant Biol. 60 (2009) 67-91.
86. A.K. Gardino, D. Kern, Methods Enzymol. 423 (2007) 149-165.
87. I. Diaz-Moreno, D. Hollingworth, T.A. Frenkiel, G. Kelly, S. Martin, S. Howell, M. Garcia-Mayoral, R. Gherzi, P. Briata, A. Ramos, Nat. Struct. Mol. Biol. 16 (2009) 238-246.
88. T. Hegedus, A.W. Serohijos, N.V. Dokholyan, L. He, J.R. Riordan, J. Mol. Biol. 378 (2008) 1052-1063.
89. M. Broncel, S.C. Wagner, K. Paul, C.P. Hackenberger, B. Koksch, Org. Biomol. Chem. 8 (2010) 2575-2579.
90. J. Mendieta, M.A. Fuertes, R. Kunjishapatham, I. Santa-Maria, F.J. Moreno, C. Alonso, F. Gago, V. Munoz, J. Avila, F. Hernandez, Biochim. Biophys. Acta. 1721 (2005) 16-26.
91. V. Babin, C. Roland, C. Sagui, J. Chem. Phys. 128 (2008) 134101-134108.
92. A. Rodriguez, C. Oliva, M. Gonzalez, Phys. Chem. Chem. Phys. 12 (2008) 8001-8015.
93. B.L. Grigorenko, A.V. Rogov, I.A. Topol, S.K. Burt, H.M. Martinez, A.V. Nemukhin, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 7057-7061.
94. A.J. Smith, Y. Li, K.N. Houk, Org. Biomol. Chem. 7 (2009) 2716-2724.
95. M. Valiev, J. Yang, J.A. Adams, S.S. Taylor, J.H. Weare, J. Phys. Chem. B. 111 (2007) 13455-13464.
96. J. Zhang, C. Li, T. Shi, K. Chen, X. Shen, H. Jiang, PLoS ONE. 4 (2009) e6304.
97. F. Shi, S.E. Telesco, Y. Liu, R. Radhakrishnan, M.A. Lemmon, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 7692-7697.
98. K. Hunter, D. Rainbow, V. Plagnol, J.A. Todd, L.B. Peterson, L.S. Wicker, J. Immunol. 179 (2007) 8341-8349.
99. H.Y. Yang, H. Yang, R. Zhao, M.H. Lee, J. Cell. Biochem. 98 (2006) 128-138.
100. C. Temporini, E. Calleri, G. Massolini, G. Caccialanza, Mass Spectrom. Rev. 27 (2008) 207-236.
101. M. Mann, S.E. Ong, M. Gronborg, H. Steen, O.N. Jensen, A. Pandey, Trends Biotechnol. 20 (2002) 261-268.
102. M. Gronborg, T.Z. Kristiansen, A. Stensballe, J.S. Andersen, O. Ohara, M. Mann, O.N. Jensen, A. Pandey, Mol. Cell. Proteomics. 1 (2002) 517-527.
103. S.S. Jensen, M.R. Larsen, Rapid Commun. Mass Spectrom. 21 (2007) 3635-3645.
104. M.R. Larsen, T.E. Thingholm, O.N. Jensen, P. Roepstorff, T.J. Jorgensen, Mol. Cell. Proteomics. 4 (2005) 873-886.
105. T.E. Thingholm, O.N. Jensen, M.R. Larsen, Proteomics. 9 (2009) 1451-1468.
106. T.E. Thingholm, O.N. Jensen, P.J. Robinson, M.R. Larsen, Mol. Cell. Proteomics. 7 (2008) 661-671.
107. H.K. Kweon, K. Hakansson, Anal. Chem. 78 (2006) 1743-1749.
108. T.S. Nuhse, A. Stensballe, O.N. Jensen, S.C. Peck, Mol. Cell. Proteomics. 2 (2003) 1234-1243.
109. X. Zhang, J. Ye, O.N. Jensen, P. Roepstorff, Mol. Cell. Proteomics. 6 (2007) 2032-2042.
110. J. Gobom, E. Nordhoff, E. Mirgorodskaya, R. Ekman, P. Roepstorff, J. Mass Spectrom. 34 (1999) 105-116.
111. M.R. Larsen, M.E. Graham, P.J. Robinson, P. Roepstorff, Mol. Cell. Proteomics. 3 (2004) 456-465.
112. A. Shevchenko, M. Wilm, O. Vorm, M. Mann, Anal. Chem. 68 (1996) 850-858.
113. M.R. Larsen, Methods Mol. Biol. 251 (2004) 245-262.
114. G.W. Becker, Brief. Funct. Genomic. Proteomic. 7 (2008) 371-382.
115. S.E. Ong, M. Mann, Nat. Chem. Biol. 1 (2005) 252-262.
116. C. Piggee, Anal. Chem. 81 (2009) 2418-2420.
117. H.C. Hemmings Jr., Protein Kinase and Phosphatase Inhibitors, Humana Press, Totowa, NJ 1996.
118. J. Reinders, A. Sickmann, Proteomics. 5 (2005) 4052-4061.
119. K. Biemann, Biomed. Environ. Mass Spectrom. 16 (1988) 99-111.
120. P. Roepstorff, J. Fohlman, Biomed. Mass Spectrom. 11 (1984) 601.
121. H. Steen, B. Kuster, M. Mann, J. Mass Spectrom. 36 (2001) 782-790.
122. H. Steen, M. Mann, J. Am. Soc. Mass Spectrom. 13 (2002) 996-1003.
123. M.J. Schroeder, D.J. Webb, J. Shabanowitz, A.F. Horwitz, D.F. Hunt, J. Proteome Res. 4 (2005) 1832-1841.
124. R.A. Zubarev, D.M. Horn, E.K. Fridriksson, N.L. Kelleher, N.A. Kruger, M.A. Lewis, B.K. Carpenter, F.W. McLafferty, Anal. Chem. 72 (2000) 563-573.
125. A.J. Kleinnijenhuis, F. Kjeldsen, B. Kallipolitis, K.F. Haselmann, O.N. Jensen, Anal. Chem. 79 (2007) 7450-7456.
126. A. Stensballe, S. Andersen, O.N. Jensen, Proteomics. 1 (2001) 207-222.
127. J.E. Syka, J.J. Coon, M.J. Schroeder, J. Shabanowitz, D.F. Hunt, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 9528-9533.
128. J.E. Elias, S.P. Gygi, Nat. Methods. 4 (2007) 207-214.
129. J. Peng, J.E. Elias, C.C. Thoreen, L.J. Licklider, S.P. Gygi, J. Proteome Res. 2 (2003) 43-50.
130. S.A. Beausoleil, J. Villen, S.A. Gerber, J. Rush, S.P. Gygi, Nat. Biotechnol. 24 (2006) 1285-1292.
131. J.D. Hoffert, G. Wang, T. Pisitkun, R.F. Shen, M.A. Knepper, J. Proteome Res. 6 (2007) 3501-3508.
132. D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell, Electrophoresis. 20 (1999) 3551-3567.
133. E. Rodriguez-Suarez, E. Gubb, I.F. Alzueta, J.M. Falcon-Perez, A. Amorim, F. Elortza, R. Matthiesen, Proteomics. 10 (2010) 1545-1556.
134. R. Matthiesen, Proteomics. 7 (2007) 2815-2832.
135. J.R. Yates, 3rd, J.K. Eng, A.L. McCormack, D. Schieltz, Anal. Chem. 67 (1995) 1426-1436.
136. L.Y. Geer, S.P. Markey, J.A. Kowalak, L. Wagner, M. Xu, D.M. Maynard, X. Yang, W. Shi, S.H. Bryant, J. Proteome Res. 3 (2004) 958-964.
137. R. Craig, R.C. Beavis, Bioinformatics. 20 (2004) 1466-1467.
138. C.E. Bakalarski, W. Haas, N.E. Dephoure, S.P. Gygi, Anal. Bioanal. Chem. 389 (2007) 1409-1419.

139. B.M. Balgley, T. Laudeman, L. Yang, T. Song, C.S. Lee, Mol. Cell. Proteomics. 6 (2007) 1599-1608.

140. D.L. Tabb, A. Saraf, J.R. Yates, 3rd, Anal. Chem. 75 (2003) 6415-6421.

141. S. Tanner, H. Shu, A. Frank, L.C. Wang, E. Zandi, M. Mumby, P.A. Pevzner, V. Bafna, Anal. Chem. 77 (2005) 4626-4639.

142. S.E. Ong, B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, M. Mann, Mol. Cell. Proteomics. 1 (2002) 376-386.

143. S. Gruhler, I. Kratchmarova, Methods Mol. Biol. 424 (2008) 101-111.

144. B.A. Ballif, P.P. Roux, S.A. Gerber, J.P. MacKeigan, J. Blenis, S.P. Gygi, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 667-672.

145. P.L. Ross, Y.N. Huang, J.N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, D.J. Pappin, Mol. Cell. Proteomics. 3 (2004) 1154-1169.

146. E. Sachon, S. Mohammed, N. Bache, O.N. Jensen, Rapid Commun. Mass Spectrom. 20 (2006) 1127-1134.

147. Z.Y. Zhang, Biochim. Biophys. Acta. 1754 (2005) 100-107.

148. P.H. Huang, W.K. Cavenee, F.B. Furnari, F.M. White, Cell Cycle. 6 (2007) 2750-2754.

149. D.S. Kirkpatrick, S.A. Gerber, S.P. Gygi, Methods. 35 (2005) 265-273.

150. G. Wang, W.W. Wu, W. Zeng, C.L. Chou, R.F. Shen, J. Proteome Res. 5 (2006) 1214-1223.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

REVIEW ARTICLE | DOI: 10.5584/jiomics.v1i1.51

Mammalian mitochondrial proteome and its functions: current investigative techniques and future perspectives on ageing and diabetes

Elena Silvestri[1,a], Assunta Lombardi[2,a], Daniela Glinni[1], Rosalba Senese[3], Federica Cioffi[3], Antonia Lanni[3], Fernando Goglia[1], Maria Moreno*[1], Pieter de Lange*[3].

[1]Dipartimento di Scienze Biologiche ed Ambientali, Università degli Studi del Sannio, Via Port'Arsa 11, 82100 Benevento, Italy; [2]Dipartimento delle Scienze Biologiche, Sez. Fisiologia ed Igiene, Università degli Studi di Napoli "Federico II", Via Mezzocannone 8, 80134 Napoli, Italy; [3]Dipartimento di Scienze della Vita, Seconda Università degli Studi di Napoli, Via Vivaldi 43, 81100 Caserta, Italy; [a]These authors contributed equally.

Received: 31 October 2010 **Accepted:** 23 November 2010 **Available Online**: 1 December 2010

ABSTRACT

Mitochondria play important roles in cellular physiology and in various pathologies. Over the last decade, great achievements have been made in study of the whole mitochondrial proteome, subproteome, mitochondrial complexes and their molecular organization, and mitochondrial post-translational modifications in both physiological and pathological conditions. Mitochondrial proteomic technologies will be discussed with regard to their applications to the detection of proteins with a net impact on the capacity of the mitochondria to perform various tasks during ageing and in diabetes.

Keywords: Mitochondrial proteome; 2-DE • BN-PAGE; OXPHOS supercomplexes; **A**geing; **D**iabetes.

Introduction

Mitochondria are essential organelles, representing, in most eukaryotic cells, the primary site for the production of ATP via oxidative phosphorylation [1]. Besides the production of energy by cellular respiration, mitochondria are involved in intermediary metabolism and in anabolic processes such as the biosynthesis of amino acids, lipids, heme and iron-sulphur clusters, the cellular homeostasis of calcium [2], and in triggering apoptosis [3].

Subserving all these functions is the unique protein composition of mitochondria, with the vast majority of the proteins being encoded within the nucleus (synthesized upon cytosolic ribosomes and imported to submitochondrial destinations) and the remaining minority being encoded by the mitochondrion's own genome, which actually encodes only a few components of the respiratory chain complexes (thirteen in humans). Thus, the mitochondrial proteome should be viewed as a dynamic program generated by fine cross-talk between the two genomes and able to adapt to the needs of the tissue or disease state. Alterations in the mitochondrial proteome exert influences over mitochondrial homeostasis, leading to several pathologies in addition to disturbing such natural processes as development and ageing [4-6]. Mitochondrial dysfunction plays roles in the pathogenesis of a wide range of diseases that involve disordered cellular fuel metabolism and survival/death pathways, including neurodegenerative diseases, cancer, and diabetes. In recent years, it has emerged that the mitochondrial proteome differs greatly among tissues, depending on their functional requirements. Moreover, it is also clear that posttranslational processes modify both the localization and the function of mitochondrial proteins in tissue- and disease-specific ways.

For a better understanding of the complex mitochondrial functions, various systematic approaches to mitochondrial-

***Corresponding author:** Maria Moreno, Dipartimento di Scienze Biologiche ed Ambientali, Università degli Studi del Sannio, Via Port'Arsa 11, 82100 Benevento, Italy, Phone: +390824305124, Fax: +39082423013, Email Adress: moreno@unisannio.it; Pieter de Lange, Dipartimento di Scienze della Vita, Seconda Università degli Studi di Napoli, Via Vivaldi 43, 81100 Caserta, Italy, Phone: +390823274580, Fax: +390823274571, Email Adress: pieter.delange@unina2.it.

protein identification and quantification have been made within the last decade. Already, proteomic strategies, predomonantly structural ones, have increased our knowledge of the mitochondrial compartment as well as of mitochondrial function.

Indeed, the proteomes of mitochondria from yeast [7-10], mouse [11-13], human cells [14,15], Neurospora crassa [16], rat [17-19], and various plants [20] have been systematically analyzed.

Proteomic analysis is being applied to mammalian cells and tissues both to gain insights into mitochondrial function and to identify disease-associated mitochondrial proteins [21,22]. Notably, such proteomic studies have proved capable of an almost complete analysis of the metabolic-flux potential of a given tissue in a single experiment, thus providing important and novel information concerning tissue- and disease- state modifications.

This article will provide an overview (a) of the current state of our knowledge about the mammalian mitochondrial proteome, with an emphasis on mitochondrial complexes and the mitochondrial PTMs that occur in ageing and diabetes, as well as (b) of the classical and ongoing approaches used to date in mitoproteomics.

1. Main tools used to analyze mitochondrial proteins: classical and ongoing approaches.

Even though the classical proteomic techniques were not initially developed specifically for the study of mitochondria, nowadays they, together with a few others, offer great, indeed unique, opportunities for the investigation of mitochondrial function within the cell. By so doing, they provide allowing to obtain structural, comparative/quantitative, and functional information.

Thanks to the early structural proteomic studies, it was possible to enumerate the mitochondrial proteins and it is now thought that, at least in mammals, there are approximately 1000 to 2000 of them. Knowledge of the abundance and the identities of these proteins has shed light on mitochondrial complexity. However, in view of the potential impact that any local mitochondrial dysfunction can have, directly or indirectly, on intra/extra mitochondrial metabolic pathways, large amounts of new data are still being sought concerning the mitoproteomic modulations that might occur in response to a given signal, a subject of interest both in physiology (i.e., ageing) and pathology (i.e., diabetes, neurodegeneration) [23,24]. The research in this field has been aided by the development of comparative proteomic techniques. These actually share the same principles as those used in structural proteomics, with the basic and widely used approaches being two-dimensional electrophoresis and shotgun proteomics, which differ only in the means used to separate and identify proteins [25,26].

In view of: the numerous functions performed by mitochondria, their high level of compartmentalization, and the existence of multipolypeptide complexes [i.e., the five components of the oxidative phosphorylation system, the Transporter Inner Membrane (TIM) and Transporter Outer Membrane (TOM) complexes, and the mitochondrial permeability transition pore] that actually contain hydrophobic proteins in close contact with the membrane lipids, peripheral proteins, and non-protein cofactors, it is likely that a deep structural/functional study of the mitoproteome requires an appropriate combination of different tools to help compensate for the limits imposed on the analysis by each individual technique. Moreover, when working on sub-cellular compartments, as in the case of mitochondria, the purity of the preparation is another critical issue. Indeed, in the process of mitochondrial preparation, several proteins can be co-isolated (i.e. from endoplasmic reticulum or other organelle membranes in close contact with mitochondria), likely impairing quantitative and qualitative analyses and, thus altering data interpretation. Several methods have been applied in efforts to obtain pure mitochondria from tissues and cells. These comprise differential centrifugation, density gradient centrifugation with Percoll™ [27], Nycodenz [28], Metrizamide [29], or sucrose [30], free-flow electrophoresis [31], and kit-based methods [32]. Mitochondria can also be highly purified by immunoisolation (by means of mitochondria-specific antibodies), although the costs are high and large amounts of sample are lost [33]. The purity of the mitochondrial preparation --a fundamental issue above all others when performing quantitative analyses of proteins with different subcellular localizations-- can be tested by measuring marker-enzyme activities and/or by Western blotting for specific markers. Thereafter, in general terms, the protein-separation strategy, critical in determining the utility of the proteomic approach, can be based either on gel electrophoresis (i.e., two-dimensional electrophoresis, DIGE, BN-PAGE) or on gel-free/MS-based techniques (i.e., shotgun proteomics). A brief description of these methods, and of their advantages and limitations, is given below.

1.1. *Two-dimensional electrophoresis (2-DE).*

2-DE allows separation of proteins first on the basis of differences in their net charge [through a technique known as isoelectric focusing (IEF); first dimension] and secondly on differences in molecular mass [through polyacrylamide gel electrophoresis (PAGE); second dimension). 2-DE is the classical, and the most popular, technique for the profiling of complex protein mixtures obtained from biological samples. It provides the highest resolution in soluble protein separation, and indeed resolves hundreds or thousands of proteins at once. In recent years, the commercial availability of preformed immobiline gradients, with various ranges of pH and sizes, has made it possible to choose the best range of pH for optimal results, depending on the origin and specific characteristics of the sample to be analyzed. The advantages of - comprise: i) fast resolution of proteins, ii) the relatively low cost of the experiments, and iii) the possibility of directly evaluating the pI and MW of the protein(s) of interest. Unfortunately, there are also several disadvantages. Among

these are: i) difficulties in the resolution of proteins with ex-extreme pI values, ii) loss of very hydrophobic proteins, iii) absence of proteins of high and low molecular weights, iv) poor resolution of low-abundance proteins, v) large amounts of sample handling, and vi) inter-gel variability. When specifically studying mitochondrial proteomes, an additional disadvantage is that, for a broad range of pH values (3–11 NL), there a slight asymmetrical distribution of the molecules, with those of molecular weights (usually) extending from 0.1 to 15 kDa, being preferentially localized in the basic parts of the gel. Proteins in 2D-gels can be detected by several staining procedures. These can be roughly divided into the following categories: (i) organic dyes; (ii) silver stains; (iii) negative stains; and (iv) fluorescent stains. For Coomassie Blue R250 (CBB), the most popular organic dye currently in use, the detection limit and linear dynamic range are from 8 to 16 ng and 125-1000 ng, respectively, depending on the proteins [34]. Silver staining is also widely used, and it is particularly suitable when the amounts of proteins are quite small [35]. Its sensitivity is relatively high (0.5-8 ng) [36], but its dynamic range is quite narrow (e.g., 8-60 ng for the alkaline/silver diamine stain) [34]. The latter obviously limits the quantitative use of silver staining when comparing samples with widely differing amounts of proteins.

Concerning negative stains, the most sensitive method is represented by the imidazole-zinc (IZ) stain. This, while having several desirable features --such as simplicity, rapidity, reversibility, and compatibility with subsequent protein analysis [37,38]-- suffers from the drawback of its poor contrast, which makes band- or spot- excision difficult [39]. Recently, a new negative detection method has been described [40]. In this method, Eosin Y is selectively precipitated in the gel background, the same precipitate being absent from those zones where proteins are located through the formation of a stable water-soluble protein-dye complex. This staining provides high sensitivity at low cost, and has a simple, fast protocol with a detection limit of 0.5 ng of a single protein band. Moreover, when compared with IZ stain, Eosin Y dye provides a broader linear dynamic range, higher sensitivity and reproducibility, and better contrast between the protein bands or spots and the background.

Among the fluorescent stains, the fluorescence-based SYPRO Ruby stain (Molecular Probes, USA) [34] has a good detection limit (0.5-5 ng) [41] and a much larger dynamic range compared to silver staining (1-1000 ng), permitting reliable quantification. Recently, Candiano et al. [42] proposed a modified Neuhoff's colloidal Coomassie Blue G-250 stain, which they dubbed "blue silver" This has high sensitivity, approaching that of conventional silver staining, with the advantages of Coomassie blue. Moreover, 2-DE electrophoresis represents the basic platform for the systematic positively identification of posttranslational modifications (PTMs) (i.e. carrying phosphate groups or groups introduced by oxidative reactions) as well as for the analysis of different splice variants. Specifically, the PTMs can be detected in gel by means of immunoblotting using antibodies against the specific added groups, or by sensitive specific probe-targeted approaches (i.e. biotin-NM).

1.2. *DIGE.*

Mainly to avoid inter-gel variability, 2D DIGE has been developed as a multiplexing technology to enable visualization of multiple protein samples on a single 2D gel [43]. Multiplexing is ensured by pre-electrophoretic labeling of protein samples with distinct and spectrally resolvable fluorescent dyes known as CyDyes (N-hydroxysuccinimide derivatives, propyl-Cy3, and methyl-Cy5). Labeling protein samples with a CyDye is very sensitive, with a detection limit of the order of ~1 ng and a dynamic range of ~105. While two independent protein samples are labeled with Cy3 and Cy5, a third fluorescent dye (Cy2) can be used to label a pooled sample containing equal amounts of the two independent samples that are to be compared [44]. This pooled internal standard is multiplexed with the Cy3- and Cy5-labeled samples within each experimental gel-replicate, and is then used for the normalization of all spots across the gel-replicates, yielding a minimized inter-gel variation and an accurate quantification of the protein-expression differences, supported by statistical analysis. Following electrophoretic separation of multiplexed labeled samples, proteins are visualized by imaging the gel at wavelengths specific for each dye allowing the identification of protein spots and quantification within each gel-replicate. Including the internal standard on each gel used in the experiment, along with the individual biological samples, means that the abundance of each protein spot on a gel can be measured relative (i.e., as a ratio) to its corresponding spot in the internal standard present on the same gel. Ettan DIGE is a system of technologies that has been optimized to benefit fully from the advantages provided by 2D DIGE [45].

1.3. *Blue native PAGE.*

The investigation of membrane-bound protein complexes (i.e., mitochondrial respiratory chain complexes, OXPHOS) requires special tools, above all because of the high hydrophobicity of the proteins involved. These have poor focalization and cannot be efficiently separated by classical 2-DE, as stated above. Consequently, a technique called Blue Native electrophoresis (BN-PAGE) was specifically developed [46-48]. It was named BN-PAGE because of the color of the crucial compound, Coomassie Blue, which gives a charge to the protein complexes without dissociating them, thereby allowing their resolution in the native form according to their molecular mass. In the years since its development, BN-PAGE has become a very important tool, particularly in mitochondrial research. Moreover, by applying 1st dimensional BN-PAGE (under native conditions) to 2nd dimensional SDS-PAGE (where electrophoresis is performed under denaturing conditions), the individual subunit proteins of the complex can be disclosed and individually investigated. The method allows analysis of both the concentration and the

composition of protein complexes. The detergent dodecyl-b-D-maltoside, which does not dissociate protein complexes, is used to solubilize the membranes, and extraction of the protein complexes is aided by the inclusion of the zwitterionic salt aminocaprionic acid. Since this has a net charge of zero at pH 7, it does not affect electrophoresis. Moreover, Cruciat et al. [49] and Schägger and Pfeiffer [50] isolated stoichiometrically assembled respiratory chain supercomplexes from yeast and mammalian mitochondria with high yields. Functionally, the assembly of OXPHOS into respirasomes has been suggested to have a number of potential advantages versus individual complexes (e.g., substrate channeling of quinones and/or cytochrome c, sequestration of reactive intermediates such as ubisemiquinone, and stabilization of individual complexes by supramolecular assembly) [50]. Supercomplexes can be obtained by substituting, during the membrane solubilization step, dodecyl-b-D-maltoside with the mild detergent digitonin [49,50]. Among the disadvantages associated with BN-PAGE, it has to be admitted that some proteins (a) cannot be separated by this electrophoretic technique because they cannot bind Coomassie blue and (b) having a neutral or basic pI, they migrate towards the cathodes and will be lost in the running buffer. Moreover, Coomassie blue can generate micelles in combination with neutral detergents, and these may interfere with fluorimetric and catalytic assays [48]. Indeed, a new technique called high resolution clear-native electrophoresis (hrCN-PAGE) was developed to try to overcome some of the disadvantages of BN-PAGE [48,51,52]. This approach, with a separation resolution as high as that of BN-PAGE, uses non-colored mixed micelles to induce a net negative charge on the proteins. This allows the performance of fluorimetric and catalytic assays. Actually, a possible disadvantage versus BN-PAGE is that hrCN-PAGE favors the dissociation of labile proteins from protein complexes.

1.4. *Shotgun proteomics and stable isotopic labeling experiments (SILE).*

In the area of mitoproteomics, the issues concerning the limitations of the above in the detection and identification of membrane (hydrophobic) proteins, as well as of low-abundance proteins, have also had to be faced during the development of large scale, so-called "shotgun proteomics", specifically with the introduction of multidimensional protein-identification technologies [53-55]. Schematically, in these procedures a complex peptide mixture is loaded on a micro capillary column packed with two stationary phases [usually, strong cation exchange (SCX) and reversed-phase (RP)] to increase protein-separation power, and this is in-line coupled between an HPLC and a tandem mass spectrometry system to enable automated protein identification and quantification [56,57]. In recent years, to add to the early adopted techniques, further approaches in shotgun proteomics have been developed, and these are being increasingly used [recently reviewed in 57]. This field is also being enriched by new, parallel developments in mass spectrometry, HPLC systems [i.e., Ultra Performance Liquid Chromatography (UPLC) technologies] as well as in peptide fragmentation techniques (i.e., ETD) which allow implementation in protein identification. Liquid chromatography (LC), which in shotgun proteomics usually supplants gel electrophoresis, has several advantages, among which are the following: it is easily automated (high throughput) and offers the possibility of studying low-abundance and highly hydrophobic proteins, which are usually masked or absent in classical 2D gels.

In quantitative shotgun proteomics the mass spectrometric (MS) data provide not only information for protein identification, but also for the determination of relative amounts, and for characterization in terms of PTM. The steps involved in protein identification by MS (viz. ionization of the peptides, separation of the obtained ions, and detection of ion masses) vary depending on the type of instrument used. Actually, various combinations of ion sources and analyzers have been developed, and each has its own advantages and inconveniences.

Paralleling the occurrence of multiplexing technology in 2-DE, quantitative LC-MS-based approaches were developed that made use of the advantages of the differential labeling of samples, which allows (provided the peptides to be compared are present in the same mass spectrum) quantitative and reproducible comparisons to be carried out. The differential labeling methods developed so far, also known as stable-isotope-labeling experiments (SILE), comprise: isotope-coded affinity tag (ICAT); isobaric tags for relative and absolute quantification (iTRAQ), and metal-coded affinity tags (MeCATs) [for technical details, see 58-60]. In 2002, Ong et al. [61] developed an additional SILE technique, now called Stable Isotopic Labeling by Amino-acids in Culture (SILAC). In this, the protein labeling is not carried out after protein extraction, but instead is "metabolically" performed during the cell culture by in vivo incorporation of either a heavy or light amino-acid during the protein translation. In general terms, SILEs offer accuracy and reproducibility in quantification, but inevitably suffer because of their intrinsic limitations (including the limited linear dynamic range, the increased time required for and increased complexity of sample preparation, and the high cost of the labeling reagents). Furthermore, to date, apart from iTRAQ, which allows comparison of up to eight samples at one and the same time, the other methods can only compare protein changes between two or three different samples.

Further, higher dynamic ranges of quantification characterize certain other proteomic approaches, namely mass spectrometry-based label-free quantitative techniques. These, although the least accurate among the MS-based quantification strategies, have proved to be successful in a variety of studies, specifically in those on the tissue-specific mitochondrial proteome changes that occur in various animal models, including normal- and diabetic- mouse models [62]. These techniques, regardless of the type of comparison --based either on measurement of the mass

spectrometric signal intensity of the peptide precursor ions of a particular protein, or on the counting of the number of fragment spectra identifying peptides of a given protein-- do not limit the number of experiments that can be compared, and so appear to be particularly advantageous for investigations of large and global protein changes between experiments.

Arguably, leaving aside the intrinsic limitations of any individual technique, the main generic advantage of 2-DE-based, as well as of non-gel based, quantitative proteomics techniques is that in one way or another, the PTMs of the proteins can be detected. In view of the central and dynamic role that mitochondria play in cellular signal transduction and in redox and nitrosative signaling, data on PTMs (such as phosphorylation, oxidation, and nitrosation) appear challenging in mitoproteomics, and indeed have novel and important information concerning the mitochondrial changes that occur during ageing and in diabetes (see below).

1.5. *Mitochondrial protein arrays.*

As happened in genomics, in proteomics, too, a variety of protein and peptide arrays have been developed (analytical as well as functional protein microarrays). In general terms, these technologies make use of protein biochips which allow the selection of specific proteins according to some of their biochemical properties (e.g., selectivity for the surface chemistry of the chip). This has the advantage of significantly reducing the complexity of the samples to be analyzed. These approaches are now widely used in biomedical studies to perform, above all, biomarker analysis, and they have begun to be applied in the field of mitoproteomics [63]. Among the various arrays, functional protein arrays have the great advantage of providing information about biochemical activities and PTMs, as well as about protein-protein, DNA-protein, RNA-protein, and drug-protein interactions.

1.6. *Mitochondrial protein databases.*

In recent years, the large-scale profiling of mitochondrial proteomes from various tissues and cells has led to the creation of more and more new databases containing mitochondrial proteome data-sets. These include Mitoproteome [64], MitoP2 [65,66], and Human Mitochondrial Protein Database and MitoMiner [67]. MitoP2, in particular, provides a comprehensive list of mitochondrial proteins from yeast, mouse, human, Arabidopsis thaliana, and Neurospora crassa, providing information about the functional annotation of proteins, as well as their subcellular locations, and their homologs, along with literature references. The recently created MitoMiner integrates 33 sets of proteomic data from six species (human, rat, mouse, Drosophila melanogaster, Caenorhabditis elegans, and Saccharomyces cerevisiae), and thereby represents the most complete mitochondrial database to date. Importantly, these data-sets have revealed that the number of mitochondrial proteins is much greater (> 3000) than the early studies estimated to be present in the mammalian mitochondrial proteome. This raises the question as to whether more mitochondrial proteins, especially membrane proteins, may be discovered in subsequent mitochondrial proteome research with the development of ad hoc proteomic technology.

A schematic representation of the overall strategies employed in mitoproteomics is shown in Figure 1.

2. Proteomic analysis pertaining to the study of mitochondria in ageing.

Several mitochondrial processes are implicated in ageing and senescence. Besides genetic determinants, increased ROS production, accumulation of oxidized proteins, and altered control of apoptosis by mitochondria have been shown to be important factors in ageing and age-related processes [68]. Proteomic studies on mammalian models have been performed in order to clarify the molecular basis of the functional deterioration that occurs in an ageing tissue [69-77]. As stated above, the precise information gained by proteomic analyses of mitochondria depends on the specific technique employed. Up to now, analytical techniques such as BN-and hrCN-PAGE combined with 2-DE have allowed the establishment of a mitochondrial protein map covering

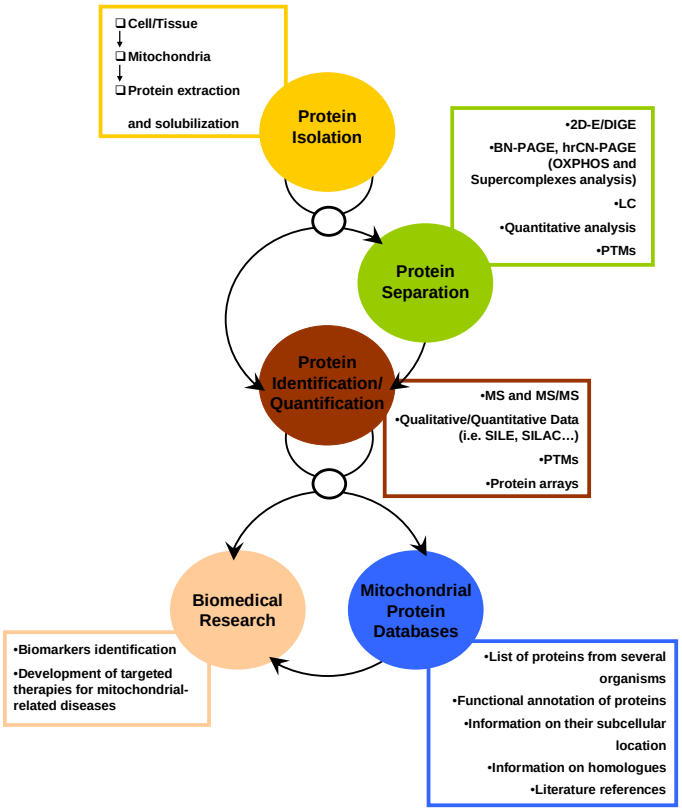


Figure 1. Main strategies and applications in mitoproteomics. The main proteomic approaches employed so far in mitochondrial studies (for mitochondrial protein separation, identification, and data analysis) are schematized, with emphasis on their mutual interrelation. For technical details and notes on both the advantages and disadvantages of the individual strategies, see text.

both the water-soluble and membrane proteins. In Table 1 reports a summary of the mitochondrial proteins so far identified through different proteomic techniques.

2.1 *Supramolecular organisation of mitochondrial respiratory chains.*

The proteomic approaches have also made possible the examination of protein redundancy, in terms of protein-protein interactions as well as PTMs. In the decade before the present one, many studies were published pointing to evidence that OXPHOS complexes do not reside as individual complexes within the inner mitochondrial membrane, but instead as large supercomplexes with a stoichiometric arrangement of complexes I, III, and IV [47,75,77-83]. Such supercomplexes are active, as shown by in-gel activity measurements within BN gels [50,84]. In mitochondria from all the tissues investigated in two studies [85,86], higher ATP synthase oligomers were separated for the first time, all of them functionally active, as displayed by in-gel ATP hydrolase activity [85].

To gain deeper insights both into ageing mechanisms and into the resulting mitoproteomic alterations, mitochondria have been studied by the BN-PAGE approach to obtain information concerning protein abundance and the supramolecular organization of OXPHOS complexes. Dencher et al. [74] reported age-modulated differences in the abundance of various proteins (see Table 1) and, age-related alterations in the oligomerization of MFoF1-ATP synthase.

The use of a combination of BN-PAGE and catalytic staining allowed detection of reduced activity in all the complexes in ageing muscle [77]. The observed reductions in the activities of respiratory complexes I, III, and V were reflections of the lower protein levels, but in contrast the reduction in complex II activity was associated with an increase in the amount of the same complex.

2.2 *Supramolecular architecture of OXPHOS complexes.*

To elucidate whether the ageing process also alters the functional/structural organization of the respiratory chain, in terms of the assembly of supercomplexes, mitochondria have been extracted using the mild detergent digitonin, since this largely retains inner mitochondrial membrane supercomplexes [50]. The major supercomplex bands in aged mitochondria were those representing the larger supercomplexes (I1 III2 IV1-4), and the majority of complex I was assembled within supercomplexes [77]. This significant increase in the supramolecular assembly of respiratory chain complexes into respirasomes might be a compensatory mechanism that, in ageing muscle, is functionally directed towards substrate-channeling and catalytic enhancement advantaging. Indeed, mitochondrial oxidative phosphorylation is more efficient in aged than in young skeletal muscle, since in old rats there is an increased respiratory control ratio (attributed principally to a reduction in the reactions able to dissipate the proton motive force not associated with ATP synthesis) [77]. This could be interpreted as a compensation for the reduced level and activity of FoF1-ATP synthase (see Figure. 2). Frenzel et al. [87] detected a profound decrease in the proportion of the supercomplexes of the respiratory chain complexes I, III2,

Table 1. Summary of the models used and of the major findings obtained by applying proteomic technologies to the study of mitochondria in ageing

Protein identified	Species, tissues	Analytical method	References
NADH dehydrogenase, cytochrome C oxidase, serum albumin precursor, casein alpha-S1, desmin, voltage-dependent anion channel 2, and adenylate kinase 2	Bovine heart	2-DE	71
3-oxoacid CoA transferase 1, ATP synthase, F1 complex α, and creatine kinase 2	mouse heart	2-DE and LC/MS	72
3-hydroxy-3-methylglutaryl-CoA synthase, acyl-CoA oxidase, enoyl-CoA hydratase, glutamine synthetase, and ornithine aminotransferase	senescence-accelerated mouse liver	2-DE and ICAT	76
NADH dehydrogenase, mitofilin, peroxiredoxin isoform III, ATPsynthase, succinate dehydrogenase, mitochondrial fission protein Fis1, succinate-coenzyme A ligase, acyl-coenzyme A dehydrogenase, porin isoform VDAC2, ubiquinol-cytochrome c reductase core I protein, and prohibitin	Rat skeletal muscle	DIGE	73
Na,K-ATP ase, HSP60, V-type ATPase, mitochondrial aconitase-2, MFoF1-ATP synthase, and complex I-IV of OXPHOS	Rat brain	BN-PAGE	74
complex I –V of OXPHOS	Rat skeletal muscle	BN-PAGE	77

and IV, as well as of the MFoF1 ATP synthase, in aged rat cerebral cortex. Complex I was observed solely in supercomplexes and was much lower in quantity in aged cortex mitochondria. An age-associated decline was observed especially in supercomplexes I1 III2, I1 III2 IV1 and I1 III2 IV2, as well as in ATP synthase, together with a pronounced increase in unbound F1 [87]. Semi-quantitative in-gel activity analysis revealed no age-dependent decline in the specific activities of the supercomplexes, favoring a compensatory mechanism for the overall decline in the abundance of respiratory supercomplexes with ageing. Contrary to the findings of Lombardi et al. [77] and Frenzel et al. [87], Gomez et al. [83] --using BN-PAGE separation of membrane proteins and LC-MS/MS analysis-- found that cardiac mitochondria display a significant age-related decline in the assemblies of supercomplexes, especially those of the highest molecular weight. These conflicting results could be due to the different model used (viz. Wistar vs. Fisher rats) or to the different methodological approaches.

2.3 *Posttranslational protein modifications.*

The nuclear transcriptional regulation of mitochondrial activity, which provides the required level of protein to accomplish a given function, is just one side of the diamond in the identification of all the mitochondrial proteins. Indeed, the dynamics of the mitochondrial proteome are manifest at multiple levels, including such PTMs as reversible phosphorylation (4-6) carbonylation, and nitrosilation. Carbonylation can be considered to be an oxidative modification that may render a protein more prone to degradation. Feng et al. [88] identified rat mitochondrial proteins that exhibited muscle type (slow- vs. fast-twitch)-dependent and age-dependent susceptibilities to carbonylation. Carbonylated mitochondrial proteins were more abundant in fast-twitch than in slow-twitch muscle. Twenty-two proteins showed significant changes in carbonylation state with age, the majority of these exhibiting increases in their amount of carbonylation. Ingenuity pathway analysis (IPA) revealed that these proteins belong to various functional classes and pathways, including cellular function and maintenance, fatty acid metabolism, and the citrate cycle. That study provided a unique catalogue of protein targets warranting further investigation because of their potential role in the muscle decline that occurs with age.

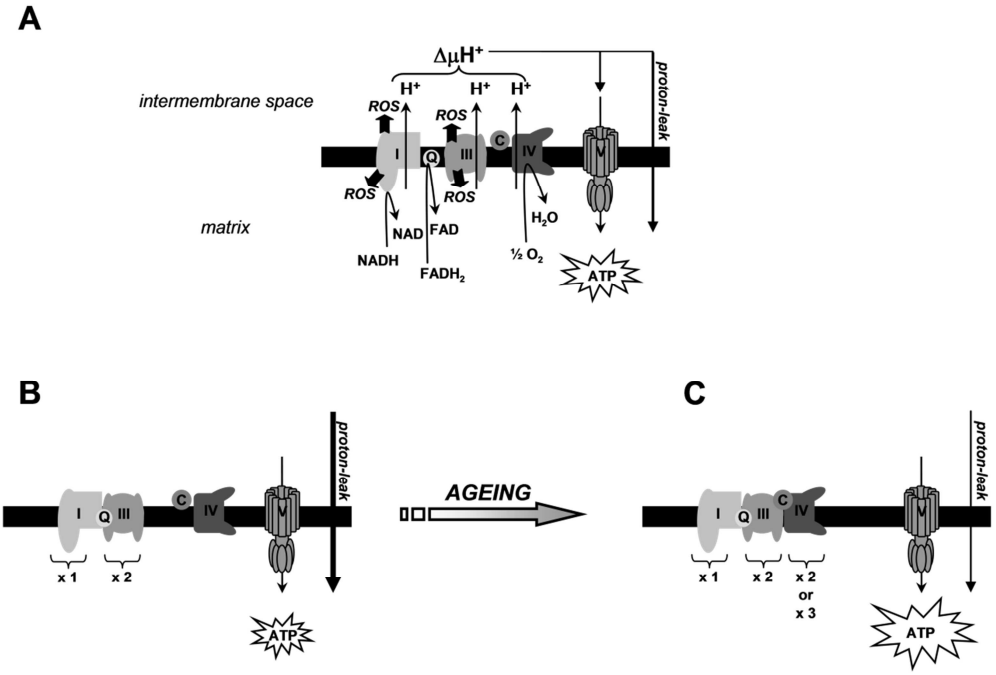


Figure 2. *Respiratory chain individual complexes (A) and their supramolecular organization in some of the main respiratory chain supercomplexes identified so far (B and C): effects of ageing in skeletal muscle and brain [according to (77, 87)].* (A) Schematic representation of mitochondrial respiratory activity. The respiratory chain (OXPHOS system) transfers electrons from reduced coenzymes (NADH and FADH2) to O2, and by pumping out H+ from the matrix into the intermembrane space, generates an electrochemical gradient, ΔμH+. This gradient provides the driving force for ATP synthesis by FoF1-ATPase (Complex V). H+ can also enter the matrix by mechanisms not coupled to ATP synthesis (generically represented as the inner membrane proton-leak). Abbreviations: I (Complex I, NADH:ubiquinone oxidoreductase), III (Complex III, ubiquinol:cytochrome c oxidoreductase), IV (complex IV, cytochrome c oxidase, COX), V (Complex V, FoF1-ATP synthase), c (cytochrome c), Q (coenzyme Q). (B and C) BN-PAGE and single particle electron microscopy has furnished evidence of specific interactions among individual protein complexes of the OXPHOS system that result in the formation of defined supramolecular structures. Various models have been suggested for these OXPHOS supercomplexes have been suggested (for review, see 78 and 108). Depending on the identity and number of the individual complexes involved, such supercomplexes may be lighter or heavier. In two different studies (77, 87), it has been demonstrated that in both skeletal muscle and brain mitochondria, the ageing process is associated with a shift towards an enrichment with heavier supercomplexes containing complex IV. This, together with a reduction in the proton-leak (at least in muscle), should represent an adaptative mechanism functionally directed towards substrate channeling and catalytic enhancement advantaging. For simplicity, dimers and supercomplexes of complex V are not represented.

Carbonylation is irreversible, however, so this modification may be of special importance in directing the affected protein along the path toward degradation.

3. Proteomic analysis pertaining to the study of mitochondria in diabetes.

Mitochondria, by virtue of their numbers or functional properties, or both, are critically involved in the pathophysiology of diabetes [89]. Comparative mitochondrial proteomics have been used to highlight the role played by mitochondria in diabetes, as well as to identify potential therapeutic markers.

Following light and heavy acrylamide labeling, 2-DE, and MS, Turko et al. [90] found that heart mitochondria from streptozotocin-induced diabetic rats displayed no alterations in the expressions of proteins from the tricarboxylic acid cycle, although an upregulation of fatty acid-oxidation proteins was detected. Other changes included down-regulations of the protein levels of creatine kinase, voltage-dependent anion channel 1, HSP60, Grp 75, and the mitochondrial-associated level of albumin but an upregulation of catalase [90]. The mitochondrial proteome has been investigated across four tissues (kidney, brain, liver and heart) from wild-type (WT) and type 1 diabetic Akita mice by means of label-free proteome expression analysis [62]. In diabetic versus WT mice, the fatty acid-oxidation proteins were less abundant in liver mitochondria, but more abundant in mitochondria from the other three tissues. In diabetes, tricarboxylic acid (TCA) cycle proteins were repressed in cardiac mitochondria, but kidney mitochondria showed a coordinated induction of TCA-cycle enzymes. Only liver mitochondria showed increased levels of OXPHOS subunits in diabetes. Mitochondrial respiration, ATP synthesis, and morphology were decreased only in cardiac mitochondria in diabetes, and these changes were accompanied by coordinated repression of OXPHOS and peroxisome proliferator–activated receptor (PPAR)-γ coactivator (PGC)-1α transcripts. These results support the idea that type 1 diabetes causes tissue-specific remodeling of the mitochondrial proteome, with a central role for mitochondrial dysfunction in diabetic cardiomyopathy [62]. Munusamy et al. [91] analyzed mitochondrial complexes by BN-PAGE, and assayed their activities in kidneys from streptozotocin-treated (diabetic) rats. They reported a reduction in Complex III, as well as a failure of this complex to assemble correctly, thus contributing to the mitochondrial oxidant production occurring in the early stages of diabetes.

Recent research findings have supported the notion that impaired mitochondrial function leads to insulin insensitivity in myocytes and adipocytes as a result either of an insufficient supply of energy or of defects in the insulin signaling pathway [92]. A comprehensive analysis of the mitochondrial proteome in the livers of type 2 diabetic rats has recently been performed by Deng et al. [93], who identified 1091 mitochondrial proteins, 228 phosphoproteins, and 355 hydroxyproteins. Their semiquantitative analysis revealed upregulation of proteins involved in beta oxidation, the TCA cycle, OXPHOS, and other bioenergetic processes, but down-regulation of proteins involved in apoptosis and oxidative stress [93]. Activation of oxidative phosphorylation was related to the overproduction of ROS, which caused oxidative stress, as also evidenced by the presence of heavily hydroxylated mitochondrial proteins [93].

Besides changes in mitochondrial protein expression, mitochondrial protein modifications play important roles in the pathology of type 2 diabetes. Protein tyrosine nitration is a common PTM occurring under conditions of oxidative stress in a number of diseases, including diabetes [94-97]. Protein tyrosine nitration alters the structure and function of proteins, and may prevent tyrosine phosphorylation [94,95,98].

Heart mitochondria from diabetic mice are prone to tyrosine nitration of proteins involved in major mitochondrial functions, such as energy production (succinyl-CoA: 3-oxoacid CoA transferase, and creatine kinase), antioxidant defense (peroxiredoxin 3), and apoptosis (voltage-dependent anion channel-1) [90]. Those findings support the general conclusion that nitration of mitochondrial proteins may result in dysfunctional mitochondria in diabetes. Kartha et al. [99] detected an increased mitochondrial protein tyrosine nitration in kidneys from high calorie- and fat diet-induced diabetic mice. Reversible phosphorylation is emerging as a central mechanism in the regulation of mitochondrial functions [100,101], and the mitochondrial phosphoproteome has been characterized in a series of studies on yeast, mouse and rat livers, porcine heart, and various plants [93,102-105]. To date, the largest data-set, published by Deng et al. [93], identified 228 different phosphoproteins and 447 phosphorylation sites in rat liver mitochondria. Hojlund et al. [106], who examined human skeletal muscle biopsies, identified eight potential protein markers for type 2 diabetes in the fasting state. The observed changes in protein expressions indicate increased cellular stress, as well as perturbations in ATP synthesis and mitochondrial metabolism, in the skeletal muscle of patients with type 2 diabetes. In particular, Hojlund et al. demonstrated that the catalytic β-subunit of ATP synthase is phosphorylated in vivo, and that the down-regulated level of the ATP synthase β -subunit phosphoisoform in diabetic muscle correlated inversely with the fasting plasma glucose level. Finally, Cui et al. [107] applied shotgun proteomics to the profiling of mitochondrial proteins, and identified potential phosphorylation sites in rat pancreatic insulinoma-1 (INS-1) beta cells. More than 800 mitochondrial proteins and 84 mitochondrial phosphoproteins were identified in that study.

4. Conclusions and future perspectives.

Mitochondrial proteomic research is expanding and will continue to expand to give further insights into the molecular mechanisms involved in both mitochondrial physiology and mitochondria-associated diseases. Proteomic approaches and

technologies developed over the last decade have provided novel insights into mitochondrial functions and have increased our knowledge of the relationship between mitochondrial dysfunction and the onset of disease. No single technique can be used for all studies of mitochondrial proteins. BN-PAGE has proved successful not only for the dissection of the inner membrane OXPHOS system, but also for the identification of the components of the outer membrane, such as those involved in protein import. Identification of PTMs, such as phosphorylation, acetylation, and nitration of mitochondrial membrane proteins, has been greatly improved by the use of affinity techniques. However, many challenges remain, largely due to the dynamic properties of these organelles. The striking differences in their composition within a given tissue and their rapid dynamics when metabolic challenges have to be faced, make mitochondria a fascinating subject for proteome-based studies in the future. An important approach could be the integration of more technologies in the fields of transcriptomics, proteomics, biochemistry, and bioinformatics, which should provide useful information concerning tissue/cell-specific mitochondrial functions.

Acknowledgements

This work was supported by GRANT-MIUR COFIN 20089SRS2X; GRANT REGIONE CAMPANIA 2008.

References

1. I.E. Scheffler, Mitochondrion 1 (2001) 3-31.
2. K. Henze, W. Martin, Nature 426 (2003) 172-6.
3. D.R. Green, J.C. Reed, Science 281 (1998) 1309-12.
4. M.R. Duchen, Mol. Aspects Med. 25 (2004) 365-451.
5. B.B. Lowell, G.I. Shulman, Science 307 (2005) 384-7.
6. D.C. Wallace, Science. 283 (1999) 1482-8.
7. H. Prokisch, C. Scharfe, D.G. 2nd Camp, W. Xiao, L. David, C. Andreoli, M.E. Monroe, R.J. Moore, M.A. Gritsenko, C. Kozany, K.K. Hixson, H.M. Mottaz, H. Zischka, M. Ueffing, Z.S. Herman, R.W. Davis, T. Meitinger, P.J. Oefner, R.D. Smith, L.M. Steinmetz, PLoS Biol. 2 (2004) e160.
8. R.P. Zahedi, A. Sickmann, A.M. Boehm, C. Winkler, N. Zufall, B. Schönfisch, B. Guiard, N. Pfanner, C. Meisinger, Molecular Cellular Biology 17 (2006) 1436-50.
9. W.K. Huh, J.V. Falvo, L.C. Gerke, A.S. Carroll, R.W. Howson, J.S. Weissman, E.K. O'Shea, Nature 425 (2003) 686-91.
10. A. Kumar, S. Agarwal, J.A. Heyman, S. Matson, M. Heidtman, S. Piccirillo, L. Umansky, A. Drawid, R. Jansen, Y. Liu, K.H. Cheung, P. Miller, M. Gerstein, G.S. Roeder, M. Snyder, Genes Dev. 16 (2002) 707-19.
11. S. Da Cruz, I. Xenarios, J. Langridge, F. Vilbois, P.A. Parone, J.C. Martinou, J Biol Chem. 278 (2003) 41566-71.
12. V.K. Mootha, J. Bunkenborg, J.V. Olsen, M. Hjerrild, J.R. Wisniewski, E. Stahl, M.S. Bolouri, H.N. Ray, S. Sihag, M. Kamal, N. Patterson, E.S. Lander, M. Mann, Cell. 115 (2003) 629-40.
13. T. Ozawa, Y. Sako, M. Sato, T. Kitamura, Y. Umezawa, Nat Biotechnol. 21 (2003) 287-93.
14. S.W. Taylor, E. Fahy, J. Murray, R.A. Capaldi, S.S. Ghosh, J Biol Chem. 278 (2003) 19587-90.
15. S.P. Gaucher, S.W. Taylor, E. Fahy, B. Zhang, D.E. Warnock, S.S. Ghosh, B.W. Gibson, J Proteome Res. 3 (2004) 495-505.
16. S. Schmitt, H. Prokisch, T. Schlunck, D.G. 2nd Camp, U. Ahting, T. Waizenegger, C. Scharfe, T. Meitinger, A. Imhof, W. Neupert, P.J. Oefner, D. Rapaport, Proteomics 6 (2006) 72-80.
17. F. Forner, E.A. Arriaga, M. Mann, J Proteome Res. 5 (2006) 3277-87.
18. D.T. Johnson, R.A. Harris, P.V. Blair, R.S. Balaban, Am J Physiol Cell Physiol. 292 (2007) C698-707.
19. D.T. Johnson, R.A. Harris, S. French, P.V. Blair, J. You, K.G. Bemis, M. Wang, R.S. Balaban, Am J Physiol Cell Physiol. 292 (2007) C689-97.
20. J.L. Heazlewood, J.S. Tonti-Filippini, A.M. Gout, D.A. Day, J. Whelan, A.H. Millar, Plant Cell. 16 (2004) 241-56.
21. S. Da Cruz, P.A. Parone, J.C. Martinou, Expert Rev Proteomics 2 (2005) 541-51.
22. A.S. Reichert, W. Neupert, Trends Genet. 20(2004) 555-62.
23. C. Ruiz-Romero, F.J. Blanco, Mol Biosyst. 5 (2009) 1130-42.
24. X. Chen, J. Li, J. Hou, Z. Xie, F. Yang, Expert Rev Proteomics. 7 (2010) 333-45.
25. T. Rabilloud, M. Chevallet, S. Luche, C.Lelong, J Proteomics (2010).
26. C.C. Wu, M.J.MacCoss, Curr Opin Mol Ther. 4 (2002) 242-50.
27. D.J. Pagliarini, S.E. Calvo, B. Chang, S.A. Sheth, S.B. Vafai, S.E. Ong, G.A. Walford, C. Sugiana, A. Boneh, W.K. Chen, D.E. Hill, M. Vidal, J.G. Evans, D.R. Thorburn, S.A. Carr, V.K.Mootha, Cell. 134 (2008) 112-23.
28. J. Li, T. Cai, P. Wu, Z. Cui, X. Chen, J. Hou, Z. Xie, P. Xue, L. Shi, P. Liu, J.R. 3rd Yates , F.Yang, Proteomics. 9 (2009) 4539-53.
29. S.W. Taylor, E. Fahy, B. Zhang, G.M. Glenn, D.E. Warnock, S. Wiley, A.N. Murphy, S.P. Gaucher, R.A. Capaldi, B.W. Gibson, S.S. Ghosh, Nat Biotechnol. 21 (2003) 281-6.
30. K. Rezaul, L. Wu, V. Mayya, S.I. Hwang, D.Han, Mol Cell Proteomics. 4 (2005) 169-81.
31. H. Zischka, G. Weber, P.J. Weber, A. Posch, R.J. Braun, D. Bühringer, U. Schneider, M. Nissum, T. Meitinger, M. Ueffing, C.Eckerskorn, Proteomics. 3 (2003) 906-16.
32. S. Hartwig, C. Feckler, S. Lehr, K. Wallbrecht, H. Wolgast, D. Müller-Wieland, J.Kotzka, Proteomics. 9 (2009) 3209-14.
33. C. Herrnstadt, W. Clevenger, S.S. Ghosh, C. Anderson, E. Fahy, S. Miller, N. Howell, R.E. Davis, Genomics. 60 (1999) 67-77.
34. K. Berggren, E. Chernokalskaya, T.H. Steinberg, C. Kemper, M.F. Lopez, Z. Diwu, R.P. Haugland, W.F. Patton, Electrophoresis. 21 (2000) 2509-21.
35. T.Rabilloud, Electrophoresis. 11 (1990) 785-94.
36. T.H. Steinberg, W.M. Lauber, K. Berggren, C. Kemper, S. Yue, W.F. Patton, Electrophoresis. 21 (2000) 497-508.
37. M.L. Ortiz, M. Calero, C. Fernandez Patron, C.F. Patron, L. Castellanos, E. Mendez, FEBS Lett. 296 (1992) 300-4.
38. L. Castellanos-Serra, W. Proenza, V. Huerta, R.L. Moritz, R.J. Simpson, Electrophoresis 20 (1999) 732-7.
39. P.L. Courchesne, R. Luethy, S.D. Patterson, Electrophoresis 18 (1997) 369-81.

40. W.T. Cong, S.Y. Hwang, L.T. Jin, H.Z. He, J.K. Choi, Electrophoresis. 31 (2010) 411-20.
41. J.C. Nishihara, K.M. Champion, Electrophoresis. 23 (2002) 2203-15.
42. G. Candiano, M. Bruschi, L. Musante, L. Santucci, G.M. Ghiggeri, B. Carnemolla, P. Orecchia, L. Zardi, P.G. Righetti, Electrophoresis. 25 (2004) 1327-33.
43. M. Unlü, M.E. Morgan, J.S. Minden, Electrophoresis. 18 (1997) 2071-7.
44. A. Alban, S.O. David, L. Bjorkesten, C. Andersson, E. Sloge, S. Lewis, I. Currie, Proteomics. 3 (2003) 36-44.
45. R. Marouga, S. David, E. Hawkins, Anal Bioanal Chem. 382 (2005) 669-78.
46. H. Schägger, G. von Jagow, Anal Biochem. 199 (1991) 223-31.
47. I. Wittig, H. Schägger, Proteomics. 9 (2009) 5214-23.
48. I. Wittig, H. Schägger, Proteomics. 8 (2008) 3974-90.
49. C.M. Cruciat, S. Brunner, F. Baumann, W. Neupert, R.A.Stuart, J. Biol. Chem. 275 (2000) 18093-8.
50. H. Schägger, K. Pfeiffer, EMBO J. 19 (2000) 1777-83.
51. I. Wittig, M. Karas, H. Schägger, Mol. Cell. Proteomics. 6 (2007) 1215-25.
52. I. Wittig, R. Carrozzo, F.M. Santorelli, H. Schägger, Electrophoresis. 28 (2007) 3811-20.
53. A.J. Link, J. Eng, D.M. Schieltz, E. Carmack, G.J. Mize, D.R. Morris, B.M. Garvik, J.R. 3rd Yates, Nat Biotechnol. 17 (1999) 676-82.
54. M.P. Washburn, D. Wolters, J.R. 3rd Yates, Nat Biotechnol. 19 (2001) 242-7.
55. D.A. Wolters, M.P. Washburn, J.R. 3rd Yates, Anal Chem. 73 (2001) 5683-90.
56. M.L. Fournier, J.M. Gilmore, S.A. Martin-Brown, M.P. Washburn, Chem. Rev. 107 (2007) 3654-86.
57. J.M. Gilmore, M.P. Washburn, Proteomics. 73 (2010) 2078-91.
58. S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, R. Aebersold, Nat Biotechnol. 17 (1999) 994-9.
59. L. DeSouza, G. Diehl, M.J. Rodrigues, J. Guo, A.D. Romaschin, T.J. Colgan, K.W. Siu, J. Proteome Res. 4 (2005) 377-86.
60. R. Ahrends, S. Pieper, A. Kühn, H. Weisshoff, M. Hamester, T. Lindemann, C. Scheler, K. Lehmann, K. Taubner, M.W. Linscheid, Mol. Cell. Proteomics 6 (2007) 1907-16.
61. S.E. Ong, B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, M. Mann, Mol. Cell. Proteomics 1 (2002) 376-86.
62. H. Bugger, D. Chen, C. Riehle, J. Soto, H.A. Theobald, X.X. Hu, B. Ganesan, B.C. Weimer, E.D. Abel, Diabetes. 58 (2009) 1986-97.
63. H.K. Nyblom, K. Thorn, M. Ahmed, P. Bergsten, Proteomics. 6 (2006) 5193-8.
64. D. Cotter, P. Guda, E. Fahy, S. Subramaniam, Nucleic Acids Res. 32 (2004) D463-7.
65. M. Elstner, C. Andreoli, U. Ahting, I. Tetko, T. Klopstock, T. Meitinger, H. Prokisch, Mol Biotechnol. 40 (2008) 306-15.
66. M. Elstner, C. Andreoli, T. Klopstock, T. Meitinger, H. Prokisch, Methods Enzymol. 457 (2009) 3-20.
67. A.C. Smith, A.J. Robinson, Mol Cell Proteomics. 8 (2009) 1324-37.
68. R.S. Balaban, S. Nemoto, T. Finkel, Cell. 120 (2005) 483-95.
69. J. Chang, H. Van Remmen, J. Cornell, A. Richardson, W.F. Ward, Mech Ageing Dev. 124 (2003) 33-41.
70. J. Chang, J.E. Cornell, H. Van Remmen, K. Hakala, W.F. Ward, A. Richardson, J. Gerontol A. Biol. Sci. Med. Sci. 62 (2007) 223-34.
71. A.N. Kiri, H.C. Tran, K.L. Drahos, W. Lan, D.K. McRorie, M.J. Horn, J Biomol Tech. 16 (2005) 371-9.
72. B. Chakravarti, M. Oseguera, N. Dalal, P. Fathy, B. Mallik, A. Raval, D.N. Chakravarti, Arch Biochem Biophys. 474 (2008) 22-31.
73. K. O'Connell, K. Ohlendieck, Proteomics. 9 (2009) 5509-24.
74. N.A. Dencher, S. Goto, N.H. Reifschneider, M. Sugawa, F. Krause, Ann. N. Y. Acad. Sci. 1067 (2006) 116-9.
75. N.A. Dencher, M. Frenzel, N.H. Reifschneider, M. Sugawa, F. Krause, Ann. N. Y. Acad. Sci. 1100 (2007) 291-8.
76. Y. Liu, J. He, S. Ji, Q. Wang, H. Pu, T. Jiang, L. Meng, X. Yang, J. Ji, Mol Cell Proteomics 7 (2008) 1737-47.
77. A. Lombardi, E. Silvestri, F. Cioffi, R. Senese, A. Lanni, F. Goglia, P. de Lange, M. Moreno, J. Proteomics 72 (2009) 708-21.
78. E.A. Schon, N.A. Dencher, Cell. Metab. 9 (2009) 1-3.
79. H. Seelert, D.N. Dani, S. Dante, T. Hauss, F. Krause, E. Schäfer, M. Frenzel, A. Poetsch, S. Rexroth, H.J. Schwassmann, T. Suhai, J. Vonck, N.A. Dencher, Biochim. Biophys Acta. 1787 (2009) 657-71.
80. I. Wittig, R. Carrozzo, F.M. Santorelli, H. Schägger, Biochim. Biophys. Acta. 1757 (2006) 1066-72.
81. C. Wernicke, J. Hellmann, B. Zieba, K. Kuter, K. Ossowska, M. Frenzel, N.A. Dencher, H. Rommelspacher, Pharmacol. Rep. 62 (2010) 35-53.
82. N.H. Reifschneider, S. Goto, H. Nakamoto, R. Takahashi, M. Sugawa, N.A. Dencher, F. Krause, J. Proteome Res. 5 (2006) 1117-32.
83. L.A. Gómez, J.S. Monette, J.D. Chavez, C.S. Maier, T.M. Hagen, Arch. Biochem. Biophys. 490 (2009) 30-5.
84. H. Eubel, J. Heinemeyer, S. Sunderhaus, H.P. Braun, Plant Physiol Biochem. 42 (2004) 937-42.
85. F. Krause, N.H. Reifschneider, S. Goto, N.A. Dencher, Biochem Biophys Res Commun. 8 (2005) 583-90.
86. F. Krause, C.Q. Scheckhuber, A. Werner, S. Rexroth, N.H. Reifschneider, N.A. Dencher, H.D. Osiewacz, Ann N Y Acad Sci. 1067 (2006) 106-15.
87. M. Frenzel, H. Rommelspacher, M.D. Sugawa, N.A. Dencher, Exp Gerontol. 45 (2010) 563-72.
88. J. Feng, H. Xie, D.L. Meany, L.V. Thompson, E.A. Arriaga, T.J. Griffin, J Gerontol A Biol Sci Med Sci. 63 (2008) 1137-52.
89. W.I. Sivitz, M.A. Yorek, Antioxid Redox Signal. 12 (2010) 537-77.
90. I.V. Turko, F. Murad, J Biol Chem. 278 (2003) 35844-9.
91. S. Munusamy, H. Saba, T. Mitchell, J.K. Megyesi, R.W. Brock, L.A. Macmillan-Crow, BMC Endocr Disord. 9 (2009) 1- 9.
92. I. Pagel-Langenickel, J. Bao, L.Pang, M.N. Sack, Endocr Rev. 31 (2010) 25-51.
93. W.J. Deng, S. Nie, J. Dai, J.R. Wu, R. Zeng, Mol Cell Proteomics. 9 (2010) 100-16.
94. S.A. Greenacre, H. Ischiropoulos, Free Radic Res. 34 (2001) 541-81.
95. I.V. Turko, F. Murad, Pharmacol Rev. 54 (2002) 619-34.
96. P. Rösen, P.P. Nawroth, G. King, W. Möller, H.J. Tritschler, L. Packer, Diabetes Metab Res Rev. 17 (2001) 189-212.
97. S.M. Grundy, I.J. Benjamin, G.L. Burke, A. Chait, R.H. Eckel, B.V. Howard, W. Mitch, S.C. Jr Smith, J.R. Sowers, Circula-

tion. 100 (1999) 1134-46.

98. H. Ischiropoulos, Arch Biochem Biophys. 356 (1998) 1-11.

99. G.K. Kartha, K.S. Moshal, U. Sen, I.G. Joshua, N. Tyagi, M.M. Steed, S.C. Tyagi, Acta Diabetol. 45 (2008) 75-81.

100. D.J. Pagliarini, J.E. Dixon, Trends Biochem Sci. 31 (2006) 26-34.

101. M. Hüttemann, I. Lee, L. Samavati, H. Yu, J.W. Doan, Biochim Biophys Acta. 1773 (2007) 1701-20.

102. J. Lee, Y. Xu, Y. Chen, R. Sprung, S.C. Kim, S. Xie, Y. Zhao, Mol Cell Proteomics. 6 (2007) 669-76.

103. J. Reinders, K. Wagner, R.P. Zahedi, D. Stojanovski, B. Eyrich, M. van der Laan, P. Rehling, A. Sickmann, N. Pfanner, C. Meisinger, Mol Cell Proteomics. 6 (2007) 1896-906.

104. E.S. Boja, D. Phillips, S.A. French, R.A. Harris, R.S. Balaban, J Proteome Res. 8 (2009) 4665-75.

105. J. Ito, N.L. Taylor, I. Castleden, W. Weckwerth, A.H. Millar, J.L. Heazlewood, Proteomics. 9 (2009) 4229-40.

106. K. Højlund, Z. Yi, N. Lefort, P. Langlais, B. Bowen, K. Levin, H. Beck-Nielsen, L.J. Mandarino, Diabetologia. 53 (2010) 541-51.

107. Z. Cui, J. Hou, X. Chen, J. Li, Z. Xie, P. Xue, T. Cai, P. Wu, T. Xu, F. Yang, J Proteome Res. 9 (2010) 2898-908.

108. N.V. Dudkina, S. Sunderhaus, E.J. Boekema, H.P. Braun, J Bioenerg Biomembr. 40 (2008) 419-24.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

REVIEW ARTICLE | DOI: 10.5584/jiomics.v1i1.43

The salivary microbiome assessed by a high-throughput and culture-independent approach

Vladimir Lazarevic*[1], Katrine Whiteson[1], Patrice François[1], Jacques Schrenzel[1].

[1]Genomic Research Laboratory, Geneva University Hospitals, Rue Gabrielle-Perret-Gentil 4, CH-1211 Geneva 14, Switzerland.

Received: 18 October 2010 **Accepted:** 14 December 2010 **Available Online**: 14 December 2010

ABSTRACT

The lack of cultivation of a significant fraction of bacteria found in the oral cavity means that culture-independent approaches are needed for the study of the salivary bacterial community composition and diversity. Saliva is easily obtained and could provide an alternative to blood in diagnostics, forensics, epidemiology and population studies. Our goal in this review is to put together the findings from the handful of recent studies of human salivary bacteria derived from culture-independent high-throughput sequencing of the 16S rRNA gene and look for emerging trends in the resulting larger dataset. Differences in phyla and genera abundances between studies of the salivary microbiome may be due to individual (genetic and lifestyle) variations, geographic variations and biases introduced during the experimental steps. Nevertheless, seven major phyla found in all relevant studies may be arbitrarily assigned into three major categories according to their abundance: (i) Firmicutes, Proteobacteria, Bacteroidetes and Actinobacteria were very abundant; (ii) TM7 and Fusobacteria were moderately abundant and (iii) Spirochaetes had the lowest abundance. The frequency of genera varied among the studies but many had a relatively consistent presence. Genera found in all of the four reported salivary microbiomes were generally present at a high frequency and contributed to 71-80% of all sequences in the corresponding datasets. Conversely, genera identified in only one available salivary microbiome generally showed a low abundance. Improvements in high-throughput sequencing technology will enable random metagenome fragment sequencing to become a powerful tool to study bacterial, archaeal, fungal, phage and human viral components of the salivary microbiome in parallel.

Keywords: Metagenomics; Microbiome; Microbiota; High-throughput Sequencing; Oral Bacteria; Saliva.

Human microbiota

The microorganisms harbored by the human body are organized in complex communities, called microbiota. They outnumber human cells by an order of magnitude [1] and their non-redundant gene set is on average more than 25 times larger than the host's gene complement [2]. Bacterial populations are not evenly distributed across the human body; they differ in density (Fig. 1), in taxa composition and abundance as well as in stability. It has been shown that microbiota variation between different habitats of the same subject is greater than interpersonal variation for the same habitat, whereas the smallest variation was observed within a given habitat over time [3].

The vast majority of bacteria colonizing the human body inhabit the distal part of the gastro-intestinal tract. Analyses of the intestinal microbiota revealed the existence of an individual core, representing the stable colonizers in a single subject, and also suggested that humans share some microbial species, the so-called universal core [4]. While the existence of a universal core was confirmed by analyzing a small number of individuals, recent studies of the intestinal microbiota including more than 100 subjects revealed that no single species-level 16S rDNA phylotype was found to be shared by all individuals [5, 6]. Therefore, it was suggested that the concept of a functional core microbiome defined at the level of shared genes, rather than shared species, would be more appropriate [6]. Yet, the issue remains a matter of the sequencing depth and definition. A deeper sample coverage and inclusion in the analysis of very rare phylotypes rather than only those occurring above a defined threshold is expected to increase the number of shared phylotypes. Indeed, deep sequencing of

***Corresponding author:** Vladimir Lazarevic, Telephone: +41 22 372 93 38, Fax: +41 22 372 98 30, Email Address: vladimir.lazarevic@genomic.ch.

metagenomic libraries from human fecal samples revealed 18 species shared by all of the 124 investigated subjects [2].

Alteration of the human microbiotas is associated with different pathologies [2, 7-10], although formal evidence that microbiota changes are actually causing such health disorders is lacking. New tools for studying human microbiota are bringing us closer to this kind of evidence. For example, the presence of particular gut microbiota has been shown to prevent intestinal inflammatory disease [11, 12], while other specific gut microbiota appear to trigger multiple sclerosis [13]. Bioremediation of microbial imbalances arises as an attractive therapeutic approach [14] and in some cases is already happening. For example, a severe *Clostridium difficile* gut infection which did not respond to several anti-microbial approaches was resolved with a healthy faucal transplant from the patient's spouse [15].

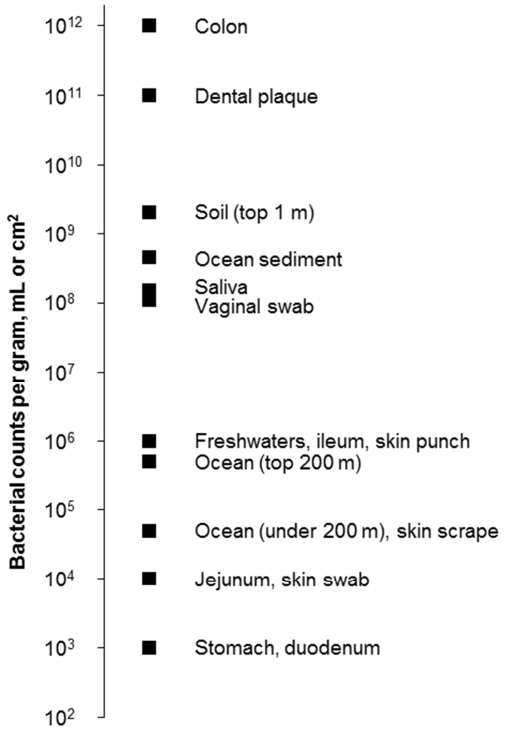


Figure 1. *Bacterial loads in different biotic and abiotic habitats.* Data were compiled from several sources in which different approaches, mainly culture-based, have been used to estimate bacterial counts [1, 16-20]. Bacterial counts on the y-axis are expressed per gram (soil, gastrointestinal and vaginal samples), per cm^3 (saliva, dental plaque, waters, ocean sediment samples) or per cm^2 (inner elbow skin samples).

Oral microbiota

Bacterial communities in the mouth have a significant impact on general health by either preventing or causing infections. Poor oral hygiene affects not only the health of the oral cavity but also the overall health of an individual by increasing risk of bacterial endocarditis and of respiratory infections [21-23]. Some oral pathologies may have a polymicrobial etiology and different types of infections appear to be associated with various mixed bacterial consortia [24]. A hypothetical causative relationship between oral microbiota profiles and oral diseases has been established [8]. Nine bacterial species were detected significantly less often in subgingival plaque samples from subjects with periodontitis compared to healthy control individuals [8]. Subjects with different forms of periodontitis can be distinguished by the frequency of 9 putative periodontal pathogenic and 15 additional species [8]. Dental caries in children are associated with a significant shift in the relative abundance of six genera in supragingival dental plaques [25].

The mouth is an entry site for passage of bacteria into the digestive tract, the respiratory tract, and the bloodstream. Since microorganisms in the mouth can translocate and colonize other parts of the body, the oral microbiota is important in the development of distant infections [26]. For example, *Streptococcus mitis* strain SF100 from the oral cavity is able to bind to human platelets with the help of phage-encoded proteins [27]. The results of a metagenomic survey of bacterial communities after the transplant of tongue bacteria to the skin suggested that different parts of the body vary in their susceptibility to colonization by oral bacteria [3].

Because of a high density and species richness, the oral cavity, including dental plaque and saliva, offers conditions that may favor genetic exchanges [28]. It has also been speculated that antibiotic-resistant oral bacteria could serve as a reservoir for the horizontal transfer of the resistance genes to other non-oral organisms that transit the oral cavity [28]. The naturally transformable bacterium *S. pneumoniae*, the major respiratory tract pathogen, may develop resistance to beta-lactam antibiotics through alterations of penicillin-binding proteins that result from genetic exchanges with commensal oral viridans streptococci [29].

More than 700 bacterial species have been identified in the human mouth and 35% of them are not yet cultivated [30]. The need to study complex oral microbiotas without culturing the bacteria prompted several research laboratories to use high-throughput sequencing of partial 16S rDNA genes amplified directly from oral bacterial communities. This approach currently provides the best compromise between sequence coverage, analytical speed and experimental costs.

Recent studies of oral microbiota using high-throughput sequencing of 16S rDNA amplicons estimate that the number of species-level phylotypes is between 540 and about 10,000 [31-33]. However, these figures were obtained using different sequencing coverage, sampling different anatomical sites and analyzing samples pooled from different numbers of individuals.

A study of the oral microbiota from three individuals, which included a mixture of samples from teeth, cheek, hard palate, tongue and saliva, supported the concept of a universal core [33]. Comparison of these oral microbiomes showed that 26.3% of distinct sequences (100%-ID phylotypes) and 47.3% of species level-phylotypes (97%-ID phylotypes) which were shared contributed to 66% and 93% of all sequence

reads of the pooled dataset, respectively. The phylotype overlap between these three oral microbiomes was significantly higher in comparison to the values reported for other body sites [5, 34], which was explained by relatively stable conditions in the human mouth [33]. In line with this observation, a survey of microbiota variation over space and time showed that the oral microbiota are less variable than those at other investigated body sites [3].

Co-occurrence analysis of bacterial taxa in data from ten individual subjects revealed genus pairs unlikely to appear together in the oral cavity [35]. By comparing the occurrence of phylotypes defined at 99% identity, evidence of possible competitive interactions between species (and/or strains) was found only within the phylum Firmicutes [35].

Salivary microbiota

Saliva contains between 10^7 to 10^9 bacteria per mL (Fig. 1), with an average value of 1.4×10^8 bacteria/mL and a higher abundance of anaerobes [18]. Since the average daily flow of saliva is between 1 and 1.5 L [36] around 10^{11} salivary bacteria may be swallowed daily. The number of bacteria attached to exfoliated oral mucosal epithelial cells was estimated to be about 2.6 times as many as those free in saliva [37]. In spite of high loads which approach those of the distal gut (Fig. 1), bacteria on the teeth and in the periodontal pockets make a small contribution to the salivary bacterial counts [38]. When 40 bacterial species were sampled on 8 oral soft tissue surfaces and studied using a DNA-DNA hybridization technique [39], salivary microbiota was found to be most similar, proportionately, to that of the dorsal and lateral surfaces of the tongue. This was confirmed by a pyrosequencing approach where the microbiota of saliva was more similar to mucosal than dental microbiotas [33].

Culture independent studies of the salivary microbiome

Several recent studies (Table 1) based on high-throughput sequencing of the preferred phylogenetic marker, the 16S rRNA gene, focused on the salivary component of the oral bacterial communities [31-33, 40, 41].

The existence of the universal core was specifically addressed in the study of the salivary microbiome from 5 individuals, each sampled at 3 time-points [41]. As expected, the size of the universal core was inversely correlated with the number of subjects sampled and the number of time points from the same individual (Fig. 2). When the three time-point samples of all individuals were taken into account, the universal core was represented by 0.3% of distinct 16S rDNA sequences and 1.9% of phylotypes defined at a 97%-ID cut-off, which corresponds to 23.3% and 37.6% of the full dataset, respectively [41]. These figures are below values reported in the study of three oral microbiomes at a single time-point [33], which had a higher sequence coverage but, besides saliva, included swabs from several oral surfaces.

The salivary microbial community appears to be relatively stable within individuals over time [41]. A large fraction of the salivary microbiome 16S sequences corresponds to the individual core. In the study which included five individuals, on average 89.0±6.1% and 77.9±7.5% of 16S rDNA sequences contributed to the individual core depending on whether the phylotypes were defined at 97 or 100% identity [41]. In the same study, time-point samples were grouped by subject using UniFrac [44]. The UniFrac metric clusters samples in terms of the phylogeny of their communities, where larger values are assigned to changes in more distant taxa. Within the same subject, samples taken at closer time intervals were not necessarily more similar than those from more distant time points, pointing to the fluctuation of some bacterial taxa between the time points. Additional analysis methods which do not give more weight to differences between more distant taxa may also be helpful in comparisons of communities, such as ANOSIM [45].

Salivary taxa abundances

We compared the abundance of phyla and genera in salivary microbiomes reported in different studies (Table 1). Comparisons were made using a pooled dataset for each study, because either samples from different individuals were not specifically labeled [31], or because the number of sequences from each individual was relatively low [40].

Differences in taxa abundance between studies may be due to individual (genetic and lifestyle) variations, geographic variations and biases introduced during the lysis procedure and PCR amplification. The 16S rRNA gene fragments were obtained by PCR amplification using primer pairs that differed between studies. Nevertheless, seven major phyla found in all studies may be arbitrarily assigned into three major categories according to their abundance (Table 1): (i) Firmicutes, Proteobacteria, Bacteroidetes and Actinobacteria were very abundant (>6%); TM7 and Fusobacteria were moderately abundant (0.5-3.7%), and (iii) Spirochaetes had the lowest abundance (0.022-0.34%). In addition, very low abundance phyla were detected in some studies. For instance, many "cyanobacterial" sequences found at very low frequency may correspond to plant chloroplasts. They are likely transient colonizers linked to food intake or exposure to airborne pollen [3]. Members of 11 other very low-abundance phyla (<0.06%) were identified sporadically. The average abundances for the top 7 phyla of the four pooled salivary microbiomes [31, 33, 40, 41] were ranked in the same order as those found in a mixture of saliva and dental plaque specimens from 10 individuals [35]. Of all bacterial phyla previously found in the oral cavity [35, 46], only Chlamydiae and OD2 were not identified in the four salivary microbiomes investigated by high-throughput sequencing of the 16S rRNA gene. The possible reasons for this are low prevalence and/or low counts of members of these two phyla in saliva as well as biases introduced during PCR amplification.

Although the frequency of genera varied among the studies, many had a relatively consistent presence (Fig. 3A). Genera found in all of the four reported salivary microbiomes were generally present at a high frequency and contributed to 71-

Table 1. Comparison of phyla abundance in salivary microbiomes

Reference	[41]	[33]	[31]	[40]
Number of subjects	5	3	71	120
Remark	3 time-points each			10 locations
16S rDNA region amplified/sequenced	V1-3/V3	V5-6	V6	V4-5
Lysis procedure	Proteinase K/ Tween-20	0.1 mm zirconium beads/phenol	0.1 mm zirconium-slilica beads/phenol	Proteinase K/SDS
Sequences analyzed	31,169	18,182[1]	73,485	14,115
Sequencing platform	GS FLX	GS FLX	GS20	Sanger
Taxonomy analysis	RDP Classifier[2]	GAST/RDP	GAST/RDP	SeqMatch/RDP
Firmicutes	53.7	47.6	40.7	37.7
Proteobacteria	20.5	16.4	21	28.6
Actinobacteria	10.4	22.9	6.3	7.0
Bacteroidetes	9.7	10.4	27.2	20.4
TM7	3.4	1.4	1.9	0.51
Fusobacteria	1.69	1.1	2.9	3.7
Spirochaetes	0.28	0.022	0.2	0.34
Bacteria	0.109	0.1	0.2	0.4
Cyanobacteria	0.064	0.049	0.02	
SR1	0.045		0.014	
Tenericutes	0.016			0.06
Synergistes				0.04
Acidobacteria			0.049	
Planctomycetes			0.018	
Nitrospira			0.0039	
Chloroflexi			0.0039	
Deinococcus-Thermus			0.0023	
Thermotogae			0.0016	
OP11			0.0016	
OD1			0.0016	

80% of all sequences in the corresponding datasets. As expected, genera identified in only one salivary microbiome generally showed a low abundance. When the average and median frequency were calculated for all of the genera shared by the given number of microbiomes (1 to 4), a positive correlation was found (Fig. 3B).

Genus *Atopobium* showed the highest variation in abundance across the studies. While the fourth most abundant genus in one [41], its members were not identified in another study [31]. Inspection of the microbiomes of the five individuals in our previous study [41] revealed significant interpersonal variation: the average frequencies of *Atopobium* based on three time-points ranged from 0.23±0.21 to 14.2±3.6. Similarly, genera *Enterobacter* and *Serratia* showed a significant interindividual variation but also a geographic patterning [40]. Both genera had a relative abundance of >4% in the salivary microbiome study reported by Nasidze *et al.* [40]. However, in three other pooled microbiome datasets, they were either absent (*Serratia*) or found only once (*Enterobacter*) at more than a 100-fold lower frequency. The observed differences in *Serratia* and *Enterobacter* frequency across studies may be due to yet-unidentified cultural and/or environmental factors [40].

Studies of the salivary microbiome based on culture-

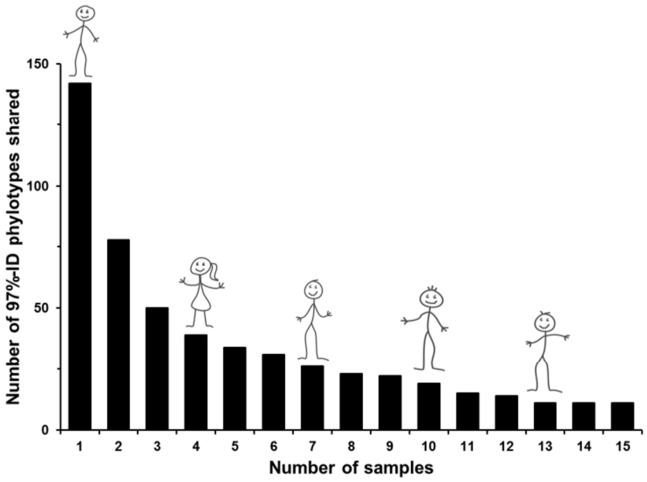


Figure 2. *Decrease in the number of species-level phylotypes shared as a function of the number of samples compared.* The bar with a stick figure above corresponds to the first time point of an individual; it is followed (to the right) by bars corresponding to the second and third time point of the same individual. The figure is based on data from our previous study [41].

independent high-throughput sequencing of partial 16S rDNA amplicons revealed genera and even higher-level taxa up to the phylum level that were not previously found in the human mouth and were not listed in the Human Oral Microbiome Database [46]. For instance, members of the class Sphingobacteria were found in saliva and/or on oral surfaces of some subjects [33, 41]. The 16S rDNA sequences corresponding to Sphingobacteria were found at a relatively low abundance (<0.11%), except in one saliva sample [41] where they represented 3.7% of all reads. Interestingly, in other saliva samples taken from the same subject at different time points within a one-month interval, Sphingobacteria were weakly present. This example provides evidence of fluctuations in oral bacterial taxa resulting in a very high relative abundance of otherwise apparently rare bacteria. Although the impact of very rare bacteria on the physiology of a bacterial community and their interaction with the host organism are not expected to be significant, these rare organisms may potentially become very abundant [41] and therefore should not be neglected a priori. Rare bacteria may also be useful as markers that are unique to individuals, as evidenced by improvements in separation of bacterial populations from individuals using a 100% sequence identity cut-off, as opposed to the standard 97% or lower cut-offs [41].

While the fraction of 16S rDNA sequences that could not be confidently placed at the phylum level is relatively small in available salivary microbiomes (Table 1), the sequences that could not be assigned at the genus level are significantly more abundant. They represent between 9.6% and 12.5% of the total number of sequences in relevant microbiome datasets [31, 33, 41]. The unassigned sequences may correspond to uncharacterized bacterial lineages confined or not to the oral niche.

Concluding remarks and outlook

After decades of culture-based investigation, culture-independent molecular techniques provide a new way to characterize salivary microbial communities. So far, these new approaches have been mainly focused on healthy adults. It will greatly benefit the field to conduct long-term surveys of a larger number of subjects in order to provide insight into the impact of different factors such as age, gender, smoking, dietary habits, oral hygiene, geographic location, use of antimicrobial compounds, immune status, underlying illness and human genetic composition on the salivary microbiota profile. Because of its relative stability, salivary microbiota may be potentially applied as an alternative or complementary approach in forensics for person identification, as recently proposed for skin bacterial communities [3].

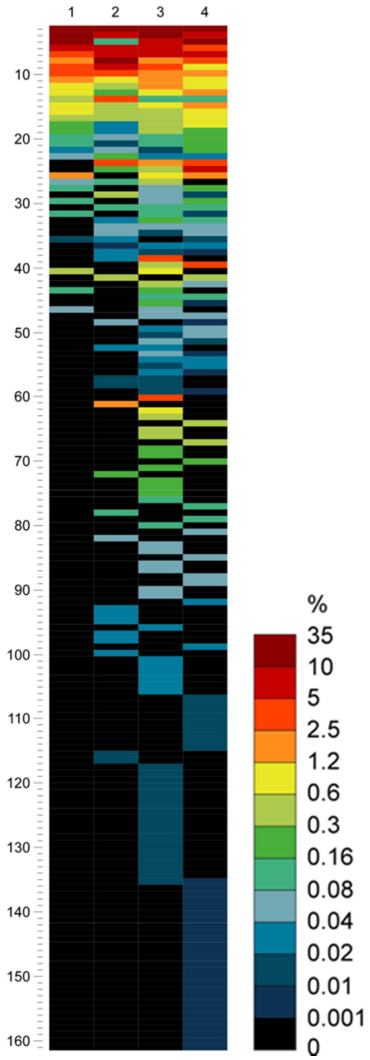


Figure 3A. *Relative abundance of bacterial genera across four (pooled) microbiomes.* Rows 1 to 161 correspond to genera listed in Supplementary Table 1 which also contains other relevant details. Genera are ranked first by the decreasing number of studies in which they were identified, then by the average frequency. The abundance is indicated according to the color scale at the right. Columns correspond to different studies as follows: 1, [31]; 2, [33]; 3, [40]; 4, [41].

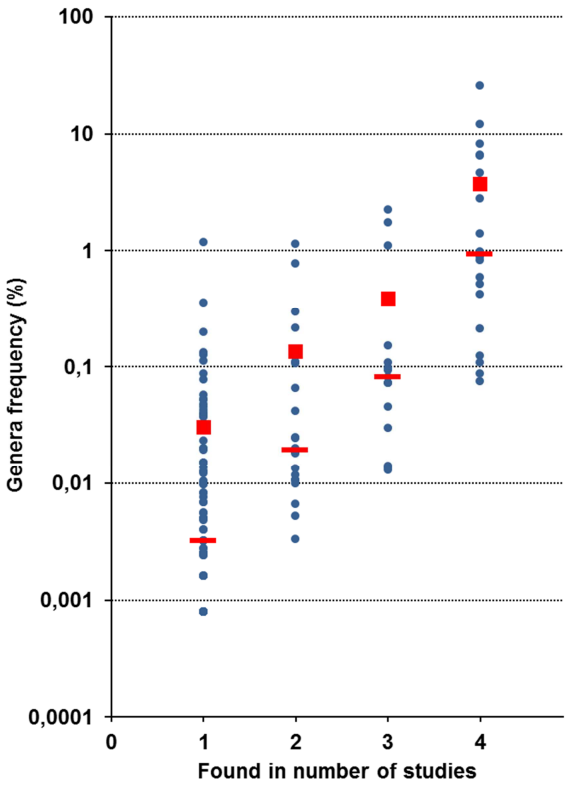


Figure 3B. *Relative abundance of bacterial genera across four (pooled) microbiomes.* Genera frequency as a function of number of studies in which they were detected. Blue circle, frequency of individual genera; red square, average genera frequency; red line, median genera frequency.

The salivary microbiome is a promising clinical diagnostic indicator of oral cancer, periodontitis [8-10] and possibly other diseases. The presence of specific pathogens and/or of a disturbed oral bacterial community might indicate the disease before symptoms are evident and may have clinical applications. Taxa which were proposed to be associated with periodontal disease before the use of high-throughput sequencing [8] are now found in healthy mouths of some individuals but often with lower abundance [35].

In addition, the human salivary microbiome may provide insights into human population structures and migrations [40], and studies with greater sequencing depth are an important background to establish what is healthy for different populations. Saliva is trivial to obtain, and can replace blood in some epidemiological studies and in diagnostics. A simple, scalable, non-invasive and cost-effective collection of saliva samples is expected to considerably increase the response rate in epidemiologic studies [47].

The fact that new taxa have been identified in each new study of the salivary microbiome and that the same genera have not been uniformly considered as universal core members across different studies, show that the our knowledge of the salivary microbiota is still developing. Therefore, larger-scale high-throughput approaches involving many time-points are required to better define the individual and universal core. Such studies, including relevant metadata records, may also allow us to understand if there are certain community structures that are more common than others and whether the oral microbial communities of healthy people can shift between alternate structures as shown for vaginal microbiota [48]. Analysis of 16S sequence fragments from oral samples taken from 300 individuals as part of the Human Microbiome Project by the National Institutes of Health, just becoming available online in 2010, may begin to establish these baselines [49].

Technological advancements in next-generation sequencing, including longer reads and paired-end strategy, may routinely provide full length 16S rDNA gene amplicon sequences in the near future, and therefore a better taxonomic resolution. This, combined with standardized DNA extraction procedures and PCR conditions, will enable more reliable comparisons of the results from different studies.

In order to minimize the impact of sequence errors, most bacterial community analyses rely on phylotypes defined at 97% identity cut-off which corresponds to the conservative estimate of bacterial species [50]. In view of the possibility that different individuals may be preferentially colonized by different strains of the same species, studies of the human microbiomes should also include analyses based on 100%-ID phylotypes [33, 41].

Ever-increasing sequencing throughput and computational power along with reference sequence databases will enable random metagenome fragment sequencing to ultimately become a standard approach for studies of oral bacterial (and other) communities. This approach has been validated and applied in the analysis of the human gut microbiome [2, 5, 51, 52]. Although a large fraction of DNA extracted from saliva corresponds to human DNA, the high sequence coverage of the whole metagenome using available HTS platforms allows us to obtain an appreciable number of bacterial and even bacteriophage sequences (unpublished). The taxonomic assignments of such identified bacterial sequences may be inferred from the BLASTN search [53] of the NCBI microbial database [54] which currently (as of December 2010) contains sequences of 1414 bacterial genomes, 879 of which are completed. Bacteriophage sequences will become increasingly identifiable as bioinformatic tools for identifying them improve; a new version of the RAST annotation engine focusing on phage annotation [55] will allow for rapid identification of known phage genes. In addition to providing a measurement of the relative abundance of bacterial taxa and the metabolic potential of the microbiota, the random sequencing of metagenomic fragments may help identify genes conferring resistance to antibiotics.

Here we have focused on bacterial communities in saliva, and we also look forward to obtaining more information about the archaeal, fungal and viral communities in the mouth, which undoubtedly play an important role in human health and will give us another angle to learn about the bacterial communities.

Archaea have not been identified in saliva so far, however, a

methanogenic archaeon *Methanobrevibacter oralis* was detected in dental plaques from a subset of patients with periodontitis. In addition, a direct correlation between the relative abundance of archaeal 16S rDNA and the severity of disease was observed [56]. The study of the archaeal component of the human oral microbiome based on the 16S rRNA sequencing requires PCR primers different from those used for bacterial identification [57].

Pyrosequencing of the internal transcribed spacer was used to characterize fungi present in the oral cavity of 20 healthy individuals [58]. Across all the samples studied, 74 culturable and 11 non-culturable fungal genera were identified, none of which was found in all sampled subjects.

By their ability to lyse bacteria, bacteriophages may play an important role in the dynamics of the oral microbiota. Bacteriophages infecting *Enterococcus faecalis* and *Proteus mirabilis* were isolated from saliva [59, 60]. Other phages, lytic for *Lactobacillus*, *Veillonella*, *Actinomyces*, *Actinobacillus*, *Streptococcus* and Proteus have been recovered from oral material including mouth wash (reviewed in [59]).

Saliva is documented to harbor several double-stranded DNA viruses (Herpes simplex virus 1, Epstein-Barr virus, cytomegalovirus and human herpesvirus 8) whose prevalence appears higher in HIV-seropositive patients [61]. Infection by viruses may have a profound effect on bacterial community composition, giving an invasive disease a chance to thrive. Viral sequences may be potentially identified in the whole salivary metagenome should the virus load be high enough and viral particles efficiently lysed.

Already the high-throughput sequencing which has been applied to the oral microbiome has revolutionized our understanding of the microbiota which reside there, giving us useful reference data for future studies of humans under diverse circumstances.

Acknowledgements

This work was supported by a grant from the Swiss National Science Foundation 3100A0-112370/1 (to J.S.) and 3100A0-116075 (to P.F.).

References

1. W.B. Whitman, D.C. Coleman, W.J. Wiebe, Proc. Natl. Acad. Sci. U S A 95 (1998) 6578-6583.
2. J. Qin, R. Li, J. Raes, M. Arumugam, K.S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D.R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H.B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, M. Jian, Y. Zhou, Y. Li, X. Zhang, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S.D. Ehrlich, Nature 464 (2010) 59-65.
3. E.K. Costello, C.L. Lauber, M. Hamady, N. Fierer, J.I. Gordon, R. Knight, Science 326 (2009) 1694-1697.
4. E.G. Zoetendal, M. Rajilic-Stojanovic, W.M. de Vos, Gut 57 (2008) 1605-1615.
5. P.J. Turnbaugh, M. Hamady, T. Yatsunenko, B.L. Cantarel, A. Duncan, R.E. Ley, M.L. Sogin, W.J. Jones, B.A. Roe, J.P. Affourtit, M. Egholm, B. Henrissat, A.C. Heath, R. Knight, J.I. Gordon, Nature 457 (2009) 480-484.
6. P.J. Turnbaugh, C. Quince, J.J. Faith, A.C. McHardy, T. Yatsunenko, F. Niazi, J. Affourtit, M. Egholm, B. Henrissat, R. Knight, J.I. Gordon, Proc. Natl. Acad. Sci U S A 107 (2010) 7503-7508.
7. J. Kuczynski, E.K. Costello, D.R. Nemergut, J. Zaneveld, C.L. Lauber, D. Knights, O. Koren, N. Fierer, S.T. Kelley, R.E. Ley, J.I. Gordon, R. Knight, Genome Biol. 11 (2010) 210.
8. A.P. Colombo, S.K. Boches, S.L. Cotton, J.M. Goodson, R. Kent, A.D. Haffajee, S.S. Socransky, H. Hasturk, T.E. Van Dyke, F. Dewhirst, B.J. Paster, J. Periodontol. 80 (2009) 1421-1432.
9. M. Faveri, M.P. Mayer, M. Feres, L.C. de Figueiredo, F.E. Dewhirst, B.J. Paster, Oral Microbiol. Immunol. 23 (2008) 112-118.
10. D.L. Mager, A.D. Haffajee, P.M. Devlin, C.M. Norris, M.R. Posner, J.M. Goodson, J Transl. Med. 3 (2005) 27.
11. S.K. Mazmanian, J.L. Round, D.L. Kasper, Nature 453 (2008) 620-625.
12. J.L. Round, S.K. Mazmanian, Proc. Natl. Acad. Sci. U S A 107 (2010) 12204-12209.
13. Y.K. Lee, J.S. Menezes, Y. Umesaki, S.K. Mazmanian, Proc. Natl. Acad. Sci. U S A (in press).
14. D.N. Frank, N.R. Pace, Curr. Opin. Gastroenterol. 24 (2008) 4-10.
15. A. Khoruts, J. Dicksved, J.K. Jansson, M.J. Sadowsky, J. Clin. Gastroenterol. 44 (2010) 354-360.
16. E.A. Grice, H.H. Kong, G. Renaud, A.C. Young, G.G. Bouffard, R.W. Blakesley, T.G. Wolfsberg, M.L. Turner, J.A. Segre, Genome Res. 18 (2008) 1043-1050.
17. A.B. Onderdonk, G.R. Zamarchi, J.A. Walsh, R.D. Mellor, A. Munoz, E.H. Kass, Appl. Environ. Microbiol. 51 (1986) 333-339.
18. R.L. Richardson, M. Jones, J. Dent. Res. 37 (1958) 697-709.
19. J. Slots, T.E. Rams, M.A. Listgarten, Oral. Microbiol. Immunol. 3 (1988) 47-52.
20. A.P. Walker, C.J. Krepel, C.M. Gohr, C.E. Edmiston, J. Clin. Microbiol. 32 (1994) 557-558.
21. K. Raghavendran, J.M. Mylotte, F.A. Scannapieco, Periodontol. 2000 44 (2007) 164-177.
22. M. Avila, D.M. Ojcius, O. Yilmaz, DNA Cell. Biol. 28 (2009) 405-411.
23. M. Wilson, Microbial inhabitants of humans : their ecology and role in health and disease, Cambridge University Press, Cambridge, 2005.
24. H.F. Jenkinson, R.J. Lamont, Trends Microbiol. 13 (2005) 589-595.
25. Z. Ling, J. Kong, P. Jia, C. Wei, Y. Wang, Z. Pan, W. Huang, L. Li, H. Chen, C. Xiang, Microb. Ecol. 60 (2010) 677-690.
26. C.L. Munro, M.J. Grap, R.K. Elswick, Jr., J. McKinney, C.N. Sessler, R.S. Hummel, III, Am. J. Crit. Care 15 (2006) 453-460.
27. J. Mitchell, P.M. Sullam, Infect. Immun. 77 (2009) 3485-3490.
28. H.K. Kuramitsu, Crit. Rev. Oral. Biol. Med. 14 (2003) 331-344.
29. C. Sibold, J. Henrichsen, A. Konig, C. Martin, L. Chalkley, R. Hakenbeck, Mol. Microbiol. 12 (1994) 1013-1023.
30. T. Chen, W.H. Yu, J. Izard, O.V. Baranova, A. Lakshmanan, F.E. Dewhirst, Database (Oxford) 2010 (2010) baq013.
31. B.J. Keijser, E. Zaura, S.M. Huse, J.M. van der Vossen, F.H. Schuren, R.C. Montijn, J.M. ten Cate, W. Crielaard, J. Dent. Res. 87 (2008) 1016-1020.
32. V. Lazarevic, K. Whiteson, S. Huse, D. Hernandez, L. Farinelli, M. Østerås, J. Schrenzel, P. François, J. Microbiol. Methods 79

(2009) 266-271.
33. E. Zaura, B.J. Keijser, S.M. Huse, W. Crielaard, BMC Microbiol. 9 (2009) 259.
34. N. Fierer, C.L. Lauber, N. Zhou, D. McDonald, E.K. Costello, R. Knight, Proc. Natl. Acad. Sci. U S A 107 (2010) 6477-6481.
35. E.M. Bik, C.D. Long, G.C. Armitage, P. Loomer, J. Emerson, E.F. Mongodin, K.E. Nelson, S.R. Gill, C.M. Fraser-Liggett, D.A. Relman, ISME J 4 962-974.
36. S.P. Humphrey, R.T. Williamson, J. Prosthet. Dent. 85 (2001) 162-169.
37. C. Dawes, Arch. Oral. Biol. 48 (2003) 329-336.
38. R.J. Gibbons, S.S. Socransky, W.C. Dearaujo, J. Vanhoute, Arch. Oral. Biol. 9 (1964) 365-370.
39. D.L. Mager, L.A. Ximenez-Fyvie, A.D. Haffajee, S.S. Socransky, J. Clin. Periodontol. 30 (2003) 644-654.
40. I. Nasidze, J. Li, D. Quinque, K. Tang, M. Stoneking, Genome Res. 19 (2009) 636-643.
41. V. Lazarevic, K. Whiteson, D. Hernandez, P. François, J. Schrenzel, BMC Genomics 11 (2010) 523.
42. Q. Wang, G.M. Garrity, J.M. Tiedje, J.R. Cole, Appl. Environ. Microbiol. 73 (2007) 5261-5267.
43. F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E.M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, R.A. Edwards, BMC Bioinformatics 9 (2008) 386.
44. C. Lozupone, M. Hamady, R. Knight, BMC Bioinformatics 7 (2006) 371.
45. K.R. Clarke, Austr. J. Ecol. 18 (1993) 117-143.
46. F.E. Dewhirst, T. Chen, J. Izard, B.J. Paster, A.C. Tanner, W.H. Yu, A. Lakshmanan, W.G. Wade, J. Bacteriol. 192 (2010) 5002-5017.
47. T.V. Hansen, M.K. Simonsen, F.C. Nielsen, Y.A. Hundrup, Cancer Epidemiol. Biomarkers Prev. 16 (2007) 2072-2076.
48. J. Ravel, P. Gajer, Z. Abdo, G.M. Schneider, S.S. Koenig, S.L. McCulle, S. Karlebach, R. Gorle, J. Russell, C.O. Tacket, R.M. Brotman, C.C. Davis, K. Ault, L. Peralta, L.J. Forney, Proc. Natl. Acad. Sci. U S A.
49. http://nihroadmap.nih.gov/hmp.
50. E. Stackebrandt, B.M. Goebel, Int. J. Syst. Bacteriol. 44 (1994) 846-849.
51. K. Kurokawa, T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, H. Morita, V.K. Sharma, T.P. Srivastava, T.D. Taylor, H. Noguchi, H. Mori, Y. Ogura, D.S. Ehrlich, K. Itoh, T. Takagi, Y. Sakaki, T. Hayashi, M. Hattori, DNA Res. 14 (2007) 169-181.
52. S.R. Gill, M. Pop, R.T. Deboy, P.B. Eckburg, P.J. Turnbaugh, B.S. Samuel, J.I. Gordon, D.A. Relman, C.M. Fraser-Liggett, K.E. Nelson, Science 312 (2006) 1355-1359.
53. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, J. Mol. Biol. 215 (1990) 403-410.
54. http://www.ncbi.nlm.nih.gov.
55. www.phantome.org.
56. P.W. Lepp, M.M. Brinig, C.C. Ouverney, K. Palm, G.C. Armitage, D.A. Relman, Proc. Natl. Acad. Sci. U S A 101 (2004) 6176-6181.
57. G.C. Baker, J.J. Smith, D.A. Cowan, J. Microbiol. Methods 55 (2003) 541-555.
58. M.A. Ghannoum, R.J. Jurevic, P.K. Mukherjee, F. Cui, M. Sikaroodi, A. Naqvi, P.M. Gillevet, PLoS Pathog. 6 e1000713.
59. G. Hitch, J. Pratten, P.W. Taylor, Lett. Appl. Microbiol. 39 (2004) 215-219.
60. G. Bachrach, M. Leizerovici-Zigmond, A. Zlotkin, R. Naor, D. Steinberg, Lett. Appl. Microbiol. 36 (2003) 50-53.
61. C.S. Miller, J.R. Berger, Y. Mootoor, S.A. Avdiushko, H. Zhu, R.J. Kryscio, J. Clin. Microbiol. 44 (2006) 2409-2415.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.19

Proteomic study in an *Escherichia coli* strain from seagulls of the Berlengas Natural Reserve of Portugal

Luís Pinto[1,2,3,4], Patrícia Poeta[3,4], Hajer Radhouani[1,2,3,4], Céline Coelho[1,2,3,4], Carlos Carvalho[1,2], Jorge Rodrigues[3,4], Carmen Torres[5], Rui Vitorino[6], Pedro Domingues[6], Gilberto Igrejas*[1,2].

[1]Department of Genetics and Biotechnology, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal; [2] Institute for Biotechnology and Bioengineering, Centre of Genetics and Biotechnology, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal; [3] Centre of Studies of Animal and Veterinary Sciences, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal; [4] Veterinary Science Department, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal; [5] Biochemistry and Molecular Biology Area, University of Rioja, Logroño, Spain; [6] Chemistry Department, University of Aveiro, Aveiro, Portugal.

Received: 8 May 2010 **Accepted:** 17 May 2010 **Available Online**: 25 June 2010

ABSTRACT

The increasing bacterial resistance among common pathogens is threatening the effectiveness of several antibiotics. This represents a serious public health problem as such bacterial strains have already been detected in domestic, wild-life animals and humans. Using *Escherichia coli* as a model organism, we applied a proteomic approach to the topic of antimicrobial resistance. In order to identify and characterize the proteome of extended-spectrum β-lactamase (ESBL) type TEM-52 producing-*Escherichia coli* strain of a faecal sample taken from Yellow-legged seagulls (*Larus cachinnans*) a bidimensional electrophoresis (2-DE) technique with an isoelectric focusing followed by a SDS-PAGE, was used. Eighty seven individualized protein spots were identified. All were suitable for peptide mass fingerprinting by a mass spectrometric technique (MALDI/TOF MS). Their identification was carried out by searching appropriate bioinformatic databases. All proteins were related to *E. coli* strains. Detection of proteins related to several *E. coli* strains linked with virulent and enterohaemorrhagic consequences in ESBL producing-*E. coli* isolates of seagull samples raises the question of how such similarities arise bearing in mind these remarkably different microbial ecosystems.

Figure 1. Seagulls (*Larus cachinnans*) in Berlengas archipelago natural reserve.

Keywords: Antibiotic Resistance; ESBL; *Escherichia coli*; Proteome; Wild Animals.

1. Introduction

Proteome analysis of innumerable biological organisms represents a major challenge for the post-genomic era and constitutes an abundant source of biological information [1]. In analogy to the genome, the proteome is used to describe a set of proteins expressed by a certain organism, under defined conditions, serving as a unique and informative readout of both its phenotypic state, which results from cell responses to physiological and environmental perturbations, and genomic information reflected in the amino acid sequences of expressed proteins. The main concern of proteomic resides, then, in the identification of proteins involved in particular cellular processes or presenting altered expression profiles as

***Corresponding author :** Dr.Gilberto Igrejas, Department of Genetics and Biotechnology, Institute for Biotechnology and Bioengineering, Centre of Genetics and Biotechnology, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal; E-mail Address: gigrejas@utad.pt.

a consequence of different physiological conditions [2].

Currently, increasing rates of bacterial resistance exist among common pathogens that are threatening the effectiveness of the most powerful antimicrobials. The emergence and spread of antibiotic-resistant Gram-negative pathogens, such as *Escherichia coli*, can lead to serious public health issues for humans and animals [3,4]. *E. coli*, a very well characterized prokaryote, has served as a model organism for several biological and biotechnological studies increasingly so since the completion of the *E. coli* genome-sequencing project [5]. The availability of complete genome sequence databases therefore facilitates the proteomic analysis of *E. coli* using MS (Mass Spectrometry). The most common resistance mechanism in *E. coli* and other Enterobacteriaceae is through the production of β-lactamases – enzymes which inactivate certain β-lactam antibiotics [6]. Extended-spectrum β-lactamases (ESBLs) are considered a major mechanism of resistance to β-lactam antibiotics that include broad-spectrum cephalosporins and azthereonam [7,8]. In fact, nowadays, ESBL-containing bacteria represent a major threat to the human community and for hospital patients causing several outbreaks and becoming endemic in many hospitals around the world [9]. The use of antibiotics in food-producing animals could lead to selection of antibiotic-resistant bacteria that can be transferred to humans through the consumption or handling of foods of animal origin. In recent studies it was demonstrated that antimicrobial resistance in foodborne bacteria may result in prolonged illness and elevated rates of bacteremia, hospitalization and death [10]. In fact, pathogenic strains of *E. coli* can cause sepsis and infections of the nervous system, digestive and urinary tracts [11] and their presence in animals in contact with humans becomes a matter of great concern.

The traditional method of measuring protein expression is by two-dimensional gel electrophoresis (2-DE) which, combined with the mass spectrometry (MS) of protein spots, makes up one of the most prevalent techniques used in proteomics, capable of generating a portrait of the global protein expression in a given sample [12,13,14]. Two-dimensional gel electrophoresis is a long established technology by which proteins are separated according to their isoelectric point (pI) and their molecular weight (Mr) [15]. The high-throughput identification of proteins excised from 2-DE gels was demonstrated by Shevchenko and colleagues (1996), who used this procedure to identify proteins expressed in yeast, following the resolution of a cell lysate by 2-D PAGE [16]. MS-based proteomics normally involves large-scale identification, quantification, and characterization of proteins at various levels resulting in an important tool for the analysis of biological systems and the exploration of complex protein functions and interactions. The most common method of protein identification is the analysis of peptide masses resulting from enzymatic digestion (e.g. by trypsin) of proteins resolved by and excised from 2-DE, by using MALDI-TOF MS [16]. In order to make sequence data derived from several bacterial genomes accessible via public databases, proteome data are characterized by diverse data types and are stored in proprietary databases located worldwide [17,18], becoming a reliable source of information for proteomic studies.

The evaluation of protein profiles in response to various stress mechanisms, such as sensitivity to antibiotics or modifications related to antibiotic resistance, could represent a valid and integrating approach for the development of new therapeutic strategies. In the current study, a proteomic evaluation of an *E. coli* isolate (GV5), recovered from a faecal sample of a Yellow-legged seagull (*Larus cachinnans*) from Berlengas Natural Reserve of Portugal and carrying an ESBL TEM-52, was performed by 2-DE and subsequent protein identification by MALDI-TOF MS.

2. Material and methods

2.1 *Cell culture and purification of E. coli*

E. coli strains were obtained from faecal samples of Yellow-legged seagulls (Figure 1), randomly recovered in the beaches of Berlengas Islands National Reserve of Portugal, located at the coast of Peniche (GPS coordinates: N 39º 24' 51,77'', W 9º 30' 33,67''). They were previously characterized for antibiotic resistant genes [19]. One of these *E. coli* strains, GV5, was included in this study. This strain was proven to be a β-lactamase TEM-52 producer [20].

2.2 *Protein extraction*

Exponentially growing cells (15 mL) were harvested by centrifugation (3 min, 10,000 xg, 4ºC) and re-suspended in PBS (4 mL) at room temperature, followed by a second centrifugation and re-suspention with SDS+Tris solution (0.2 mL) [12]. Cell disruption was performed by sonication (3× 10 s, 4 °C at 100 W); cell debris was removed by centrifugation (14,000 xg, 30 min at 4ºC). The protein concentration was assayed using a 2D Quant kit (GE Healthcare).

2.3 *One-dimensional electrophoresis and staining*

One-dimensional electrophoresis was conducted with SDS-polyacrylamide gels (T=12.52%, C=0.97%) in a HoeferTM SE 600 Ruby® (Amersham Biosciences) unit, following Laemmli [21] with some specific modifications [22]. Gels were stained during 24 hours in Coomasse Brilliant Blue R-250 and washed in water overnight. It was then fixed in trichloroacetic acid (6%) for four hours and in glycerol (5%) for two hours [23].

2.4 *Two-dimensional electrophoresis and proteome analysis*

2-DE was performed according to the principles of O'Farrell (1975) but with IPG (ImmobilineTM pH Gradient) technology [21]. Protein samples of *E. coli* (GV5) were used in parallel with samples of *E. coli* C583 and C580. For IEF, precast IPG strips with linear gradient of pH 3-10 were passively rehydrated overnight (12 to 16 hours) in a reswelling tray with rehydration buffer (8M urea, 1% CHAPS, 0.4% DTT, 0.5% carrier ampholyte IPG buffer pH 3-10) at room

temperature. IPG strips were covered with DryStrip Cover Fluid (Plus One, Amersham Biosciences). Lyses buffer [9.5M urea, 1% (w/v) DTT, 2% (w/v) CHAPS, 2% (v/v) carrier ampholytes (pH 3-10) and 10 mM Pefabloc® proteinase inhibitor] was added to the *E. coli* isolates (1:1). Samples containing a total of 100 µg of protein were loaded into 13 cm IPG strips (pH 3–10 NL, Amersham Biosciences, UK) [23]. The sample solution was then applied to the previously rehydrated IPG strips pH3-10 by cup loading and then proteins were focused sequentially at 500 V for 1 h, gradient at 1000 V for 8 h, gradient at 8000 V for 3 h, and finally 8000 V during 1 h on an EttanTM IPGPhor IITM (Amersham Biosciences, Uppsala, Sweden). Seven IEF replicate runs were performed according to Görg [24] and the GE Healthcare protocol for IPG strips pH 3-10 of 13 cm, in order to obtain the optimized running conditions, resulting in a final 13 hour run. Focused IPG strips were then stored at -80ºC in plastic bags. Before running the second dimension, strips were equilibrated twice for 15 minutes in equilibration buffer (6 M urea, 30% (w/v) glycerol, 2% (w/v) SDS in 0.05 M Tris-HCl buffer (pH 8.8)). In the first equilibration 1% DTT was added to the original equilibration buffer and to the second, 4% iodoacetamide. Bromophenol blue was added to both solutions. The equilibrated IPG strips were gently rinsed with SDS electrophoresis buffer, blotted to remove excessive buffer, and then applied to 12.52% polyacrylamide gels in a HoeferTM SE 600 Ruby® (Amersham Biosciences) unit. Some modifications were introduced in the SDS-PAGE technique previously reported by Laemmli (1970), that allowed its resolution to be increased, with proper insertion of the IPG strips in the stacking gel [21,22]. After SDS-PAGE, the 2-DE gels were fixed in 40% methanol / 10% acetic acid for one hour and afterwards stained overnight in Coomassie Brilliant Blue G-250 [25]. Coomassie-stained gels were scanned on a flatbed scanner (Umax PowerLook 1100; Fremont, CA, USA), and the resulting digitized images were analyzed using Image Master 5.0 software (Amersham Biosciences; GE Healthcare).

2.5 *Protein identification by MALDI-TOF/TOF*

To increase experimental efficacy, four separate gels were analyzed originally representing three independent *E. coli* protein samples that were previously pooled together and compared. Spots of expression in all gels were manually excised from the gels and analyzed using Matrix-Assisted Laser Desorption/Ionization-Time of Flight Mass Spectrometry (MALDI-TOF). The gel pieces were washed three times with 25 mM ammonium bicarbonate /50 % ACN (acetonitrile), once with ACN and dried in a SpeedVac (Thermo Savant). 25 mL of 10 mg/mL sequence grade modified porcine trypsin (Promega) in 25 mM ammonium bicabornate was added to the dried gel pieces and the samples were incubated overnight at 37ºC. Extraction of tryptic peptides was performed by addition of 10% of formic acid (FA)/50% ACN followed by three-fold lyophilisation in a SpeedVac (Thermo Savant). Tryptic peptides were re-suspended in 10 mL of a 50% acetonitrile/0.1% formic acid solution. The samples were mixed (1:1) with a matrix consisting of a saturated solution of a-cyano-4-hydroxycinnamic acid prepared in 50% acetonitrile/ 0.1% formic acid. Aliquots of samples (0.5 µL) were spotted onto the MALDI sample target plate.

Peptide mass spectra were obtained on a MALDI-TOF/TOF mass spectrometer (4800 Proteomics Analyzer, Applied Biosystems, Europe) in the positive ion reflector mode. Spectra were obtained in the mass range between 800 and 4500 Da with ca. 1500 laser shots. For each sample spot, a data dependent acquisition method was created to select the six most intense peaks, excluding those from the matrix, trypsin autolysis, or acrylamide peaks, for subsequent MS/MS data acquisition. Mass spectra were internally calibrated with autodigest peaks of trypsin (MH+: 842.5, 2211.42 Da) allowing a mass accuracy of better than 25 ppm.

2.6 *Database search*

Spectra were processed and analyzed by the Global Protein Server Workstation (Applied Biosystems), which uses internal MASCOT software (v 2.1.04, Matrix Science, London, UK) for searching the peptide mass fingerprints and MS/MS data. Swiss-Prot nonredundant protein sequence database (Release 57 of March 2009, 428650 entries) was used for all searches under *E. coli*. The database search parameters were as follows: carbamidomethylation and propionamide of cysteine (+71Da) as a variable modification as well as oxidation of methionine (+16Da), and the allowance for up to two missed tryptic cleavages. The peptide mass tolerance was 25 ppm and fragment ion mass tolerance was 0.3 Da. Positive identifications were accepted up to 95% of confidence level.Protein identifications were considered as reliable when the MASCOT score was > 70 (MASCOT score was calculated as $-10 \times \log P$, where P is the probability that the observed match is a random event.). This is the lowest score indicated by the program as significant ($P < 0.05$) and indicated by the probability of incorrect protein identification.

3. Results and Discussion

E. coli strains included in the study were previously characterized for virulence and antibiotic resistance. The seagull's ESBL-producing *E. coli* strain GV5 presented resistance to ampicillin, cefotaxime, ceftazidime, naladixic acid, ciprofloxacine, tetracycline, streptomycin, trimethoprim-sulfametoxazole, and chloramphenicol and contained the following resistance genes: blaTEM-52 (ESBL gene encoding TEM-52) cmlA (chloramphenicol-resistance), tetA (tetracycline-resistance), sul1 and sul3 (sulfametoxazole-resistance). The intI1gene encoding the integrase of class I integrons was also identified in strain GV5, as well as the gene cassettes dfrA1 + aadA1 included in their variable region. *E. coli* GV5 was classified into the phylogenetic group D, mostly associated with virulent extraintestinal isolates [19].

In sample GV5 a total of 87 protein spots were collected and individualized in eppendorfs for their analysis using MALDI-TOF mass spectrometry. The protein spots were identified and characterized (Table 1 on supplementary ma-

terial) by correlation with bioinformatic databases (http://www.ncbi.nlm.nih.gov/). From a total of 87 protein spots, nine proteins were detected as related to *E. coli* serotype O6 and another six proteins linked with *E. coli* serotype O157:H7 (Figure 2, A and B respectively; see also Table 1). Among the proteins identified as linked to serotype O6, it is important to highlight the presence of proteins related to environmental stress conditions: the curved DNA-binding protein (spot 18), the chaperone protein (spot 35) and the GrpE protein (spot 45) [26,27,28] are responsible for the capacity of the bacteria to resist external damage (heat, osmotic and other environmental stress causes). Beta-lactamase TEM (P62593, spot 39) and dihydrofolate reductase type XV (P78218, spots 8) were related to antibiotic resistance. The trigger factor protein (or tig), on other hand, is involved in the protein exportation whereas the other proteins related to serotype O6 presented functions in important metabolic pathways essential for the bacteria. The proteins found to be related to serotype O157:H7 presented various functions, as kdsA protein (spot 60) with an important part in the external membrane and lipopolysaccharide biogenesis [27]. The curved DNA-binding protein previously identified for the O6 serotype, was also found in the O157:H7 serotype (spot 18).

The *E. coli* serotype O6 represents a heterogeneous group of bacteria which differ in the genotypic presence as well as in the phenotypic expression of virulence factors, being already detected in humans, dogs and cats [29]. On the other hand, the serotype O157:H7 is an enterohemorrhagic strain that belongs to the verotoxin-producing *E. coli* (VTEC). VTEC have emerged as food-borne pathogens related to gastroenteritis that may be complicated by hemorrhagic colitis or by the hemolytic uremic syndrome, the main cause of acute renal failure in children [30]. The transmission of these food-borne pathogens occurs through consumption of undercooked meat, unpasteurized dairy products, vegetables, water contaminated by ruminant feces as these animals can carry *E. coli* O157:H7 transiently and sporadically and pass the bacteria in their feces [30,31,32]. The proximity of the Berlengas Natural Reserve to the Continental Portuguese coastline is an important factor in considering the possible contact between seagulls and potential disease transmitters. In fact, vero cytotoxin-producing *E. coli* O157 has already been detected in the past in faecal samples from wild-birds [33] namely from gulls [34], determining probably a transmission pattern. 2-DE proteomics correlated with bioinformatic databases has already been used in analyzing and characterizing bacterial proteins, as in *E. coli* where a large number of proteins are already completely identified [5,35]. In our study a great number of proteins which were related to several functions within the cell metabolism were found, like β-lactamase TEM precursor proteins (blaT) capable of producing β-lactamases (Figure 3). The blaT proteins are prevalent in enterobacteriaceae hydrolyzing the β-lactam bond of susceptible β-lactam antibiotics like penicillins and cephalosporins [36,37]. The presence of three protein spots identified as blaT (spots 39, 40 and 79; see Table 1) confirms that our sample possesses resistance to β-lactam antibiotics. Other proteins relevant to various important and essential functions in the bacteria were found. Flagellin or FliC (spot 14) is responsible for the formation of bacterial flagella [28,38]; chaperone protein or DnaK (spots 1, 49 and 62), involved in the chromosomal DNA replication [39,40,41]; while L-asparaginase II or AspG2 (spot 53), is an important therapeutic enzyme for the treatment of leukemia [28,42].

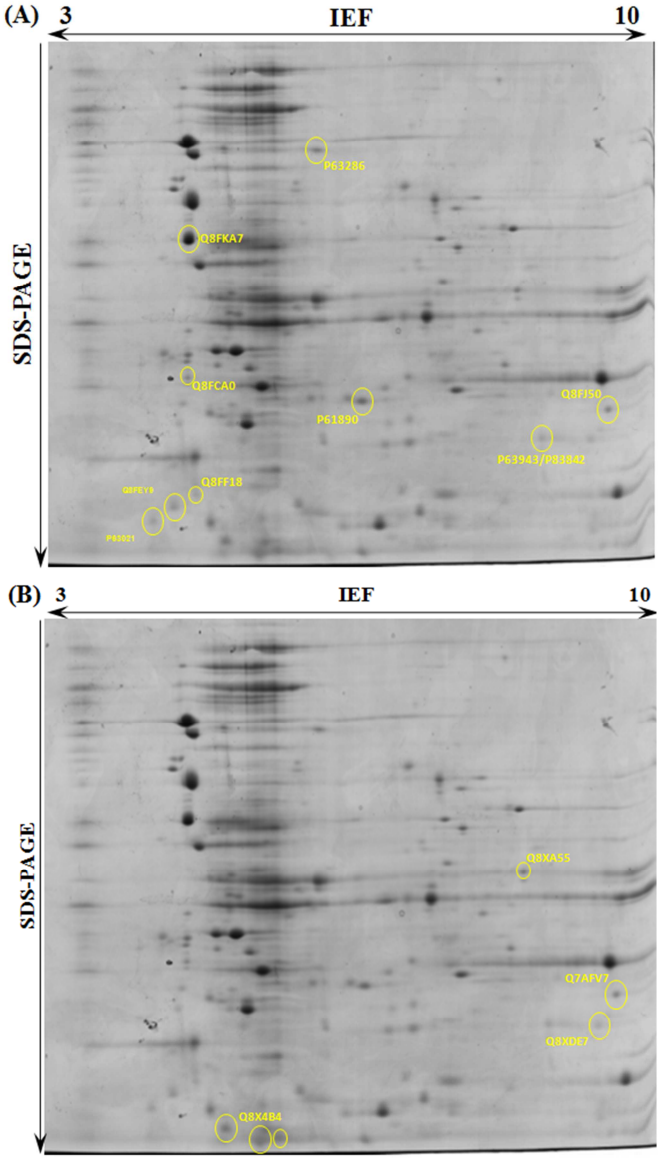


Figure 2. 2-DE gel image of *E. coli* sample GV5. (A) Accession numbers of proteins related to *E. coli* serotype O6; (B) Accession numbers of proteins related to *E. coli* serotype O157:H7.

In the case of enterohemorrhagic *E. coli* strain, four related proteins were found in *E. coli* isolates GV5 (Putative flavoprotein, Serine hydroxymethyltransferase protein and Curved-DNA binding protein). Also noticed was the detection of proteins Malate dehydrogenase and GrpE protein linked with *E. coli* enterohemorrhagic strain O157:H7 in *E. coli* virulent strain O6 in the seagull isolate.

The presence of proteins associated with such *E. coli* strains

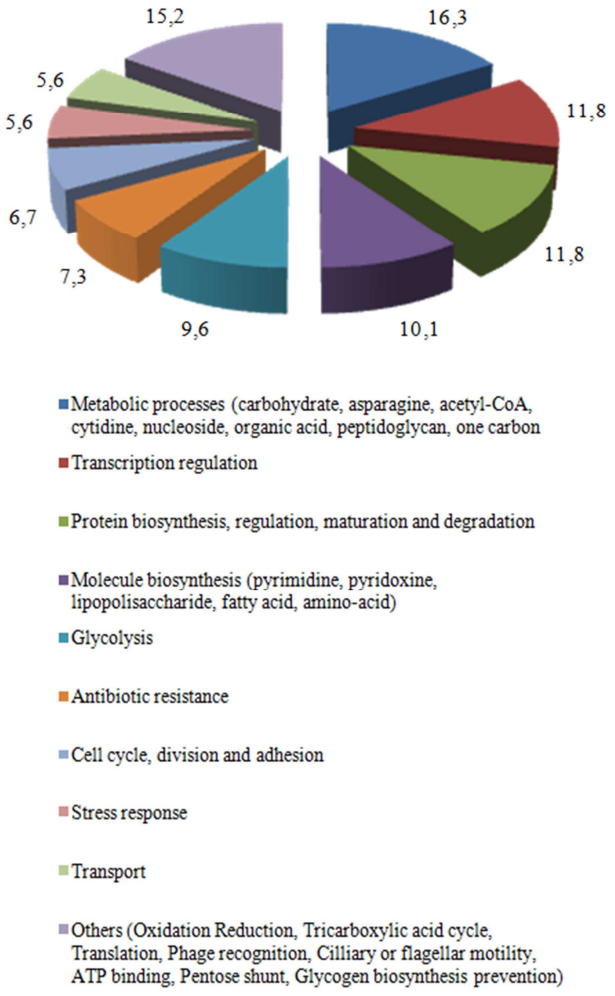


Figure 3. Distribution of the biological processes related to the protein spots found in the 2-DE gels of the *E. coli* GV5.

alongside proteins related to commensal strains in faecal samples of wild-life animals and their resistance to antimicrobial drugs represents a public health concern. The idea of antimicrobial resistance in the same extended-spectrum β-lactamase *E. coli* of very different ecosystems as microbial fauna of wild-life animals and humans with repressed imunoystems is therefore a new concern in the already problematic question of antibiotic resistance and possible transmission.

4. Concluding remarks

In this study, the elaboration of a 2-DE electrophoresis gel of an extended-spectrum β-lactamase *E. coli* strain with phenotypic and genotypic profiles indicating antimicrobial resistance allowed us to identify and characterize the proteins present. The proteome patterns obtained reveal proteins previously identified in the virulent strain *E. coli* O6 and enterohemorrhagic strain *E. coli* O157:H7. The detection of proteins related to these strains in samples of extended-spectrum β-lactamase-producing *E. coli* isolates became possible through the proteomic approach and 2-DE combined with mass spectrometry. Considering this work, it is possible to elucidate gene expression of multiresistant bacteria strains isolated from different wild ecosystems. In the future it will be important to evaluate this expression under different forms of stress. These proteins should be tested under stress conditions, for example under antibiotic pressure, in order to determine the changes in protein expression, and to test potential targets for designing new drugs to inhibit the growth of the antibiotic-resistant bacteria.

5. Supplementary material

Supplementary material regarding this manuscript is online available in the web page of JIOMICS.

http://www.jiomics.com/index.php/jio/rt/suppFiles/19/0

References

1. S.D. Patterson, RH Aebersold, Nat. Genet. 33 (2003) 311–323.
2. E.A. Panisko, T.P. Conrads, M.B. Goshe, T.D. Veenstra, Exp. Hematol. 35 (2002) 97-107.
3. D.N. Fish, M.J. Ohlinger, Crit. Care Clin. 22(2) (2006) 291-311.
4. Infectious Diseases Society of America. Bad bugs, no drugs. As antibiotic discovery stagnates ... a public health crisis brews. Alexandria, VA: Infectious Diseases Society of America, 2004.
5. M. Han, S.Y. Lee, Microbiol. Molec. Biol. Rev. 70(2) (2006) 362-439.
6. G.A. Jacoby, L.S. Munoz-Price, New England Journal of Medicine 352 (2005) 380-391.
7. K. Bush, Clin. Infect. Dis. 32 (2001) 1085-1089.
8. N. Mendonça, J. Leitão, V. Manageiro, E. Ferreira, the Antimicrobial Resistance Surveillance Program in Portugal, M. Caniça, Antimicrob. Agent. Chemother. 51(6) (2007) 1946-1955.
9. F. Perez, A. Endimiani, K.M. Hujer, R.A. Bonomo, Curr. Opin. Pharmacol. 34(11) (2004) 514-521.
10. L. Tollefson, B.E. Karp, Med. Mal. Infect. 34(11) (2004) 514-521.
11. E. Gustafsson, K. Thóren, T. Larsson, P. Davidsson, K. Karlsson, C.L. Nilsson, Rapid Commun. Mass Spectrom. 15 (2001) 428-432.
12. J.E. Celis, P. Gromov, Curr. Opin. Biotechnol. 10(1) (1999) 16-21.
13. M.L. Yarmush, A. Jayaraman, Annu. Rev. Biomed. Eng. 4 (2002) 349-373.
14. E. Dumas, M. Desvaux, C. Chambon, M. Hébraud, Proteomics 9 (2009) 3136-3155.
15. P.H. O´Farrell, J. Biol. Chem. 250 (1975) 4007-4021.
16. A. Shevchenko, O.N. Jensen, A.V. Podtelejnikow, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, A. Shevchenko, H. Boucherie, M. Mann, Proc. Acad. Sci. USA 93 (1996) 14440-14445.
17. D. Hiscock, C. Upton, Bioinformatics 16(5) (2000) 484-485.
18. J.M. Oh, S.M. Hanash, D. Teichroew, Electrophoresis 20(4-5) (1999) 766-774.
19. P. Poeta, H. Radhouani, G. Igrejas, A. Gonçalves, C. Carvalho, J. Rodrigues, L. Vinué, S. Somalo, C. Torres, Appl. Environ. Microbiol. 74 (2008) 7439-7441.
20. L. Vinué, M. Lantero, Y. Sáenz, S. Somalo,I. de Diego, F. Pérez, F. Ruiz-Larrea, M. Zarazaga, C. Torres, J. Med. Microbiol. 57 (2008) 916-920.

21. U.K. Laemmli, Nature 227 (1970) 680-685.
22. G. Igrejas, Genetic, biochemical and technological factors associated to the utilization of common wheat (Triticum aestivum L.), University of Trás-os-Montes and Alto Douro 2000, Ph.D. Thesis.
23. A. Görg, C. Obermaier, G. Boguth, A. Harder, B. Scheibe, R. Wildgruber, W. Weiss, Electrophoresis 21 (2000) 1037-1053.
24. A. Görg, A. Klaus, C. Lück, F. Weiland, W. Weiss, Technische Universität München 2007.
25. A. Görg, W. Weiss, M.J. Dunn, Proteomics 4 (2004) 3665-3685.
26. R.A. Welch, V. Burland, G. Plunkett 3rd, P. Redford, P. Roesch, D. Rasko, E.L. Buckles, S.R. Liou, A. Boutin, J. Hackett, D. Stroud, G.F. Mayhew, D.J. Rose, S. Zhou, D.C. Schwartz, N.T. Perna, H.L. Mobley, M.S. Donnenberg, F.R. Blattner, Proc. Natl. Acad. Sci. USA 99(26) (2002) 17020-17024.
27. N.T. Perna, G. Plunkett 3rd, V. Burland, B. Mau, J.D. Glasner, D.J. Rose, G.F. Mayhew, P.S. Evans, J. Gregor, H.A. Kirkpatrick, G. Pósfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E.J. Grotbeck, N.W. Davis, A. Lim, E.T. Dimalanta, K.D. Potamousis, J. Apodaca, T.S. Anantharaman, J. Lin, G. Yen, D.C. Schwartz, R.A. Welch, F.R. Blattner, Nature 409(6819) (2001) 529-533.
28. F.R. Blattner, G. Plunkett 3rd, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, J. Gregor, N.W. Davis, H.A. Kirkpatrick, M.A. Goeden, D.J. Rose, B. Mau, Y. Shao, Science 277(5331) (1997) 1453-1474.
29. J.R. Johnson, B. Johnston, C.R. Clabots, M.A. Kuskowski, E. Roberts, C. DebRoy, J. Clin. Microbiol. 46(2) (2008) 417-422.
30. J. Blanco, M. Blanco, J.E. Blanco, Mora A., E.A. González, M.I. Bernárdez, M.P. Alonso, A. Coira, A. Rodríguez, J. Rey, J.M. Alonso, M.A. Usera, Exp. Biol. Med. 228 (2003) 345-351.
31. W.C. Cray Jr., H.W. Moon, Appl. Environ. Microbiol. 61 (1995) 1586-1590.
32. J.S. Wallace, CABI Publish. (1999) 195-223.
33. G. Foster, J. Evans, H.I. Knight, A.W. Smith, G.J. Gunn, L.J. Allison, B.A. Synge, T.W. Pennycott, Appl. Environ. Microbiol. 72(3) (2006) 2265-2267.
34. [34] J.S. Wallace, T. Cheasty, K. Jones, J. Appl. Microbiol. 82(3) (1997) 399-404.
35. A. Lopez-Campistrous, P. Semchuck, L. Burke, T. Palmer-Stone, S.J. Brokx, G. Broderick, D. Bottorff, S. Bolch, J.H. Weiner, M.J. Ellison, Molec. Cell Proteom. 4 (2005) 1205-1209.
36. T. Brun, J. Péduzzi, M.M. Caniça, G. Paul, P. Névot, M. Barthélémy, R. Labia, FEMS Microbiol. Lett. 120(1-2) (1994) 111-117.
37. C. Mabilat, J. Lourençao-Vital, S. Goussard, P. Courvallin, Mol. Gen. Genet. 235(1) (1992) 113-121.
38. K. Hayashi, N. Morooka, Y. Yamamoto, K. Fujita, K. Isono, S. Choi, E. Ohtsubo, T. Baba, B.L. Wanner, H. Mori, T. Horiuchi, Mol. Syst. Biol. 2 (2006) 1-5.
39. J.C. Bardwell, E.A. Craig, Proc. Natl. Acad. Sci. USA 81(3) (1984) 848-852.
40. J.C. Bardwell, K. Tilly, E. Craig, J. King, M. Zylicz, C. Georgopoulos, J. Biol. Chem. 261(4) (1986) 1782-1785.
41. M. Pellechia, D.L. Montgomery, S.Y. Stevens, C.W. Vander Kooi, H.P. Feng, L.M. Gierasch, E.R. Zuiderweg, Nat. Struct. Biol. 7(4) (2000) 298-303.
42. C. Derst, J. Henseling, K.H. Röhm, Prot. Eng. 5(8) (1992) 785-789.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.20

Proteomic evaluation of Escherichia coli isolates from human clinical strains

Luís Pinto[1-4], Patrícia Poeta[3,4], Hajer Radhouani[1-,4], Céline Coelho[1-4], Carlos Carvalho[1,2], Jorge Rodrigues[3,4], Carmen Torres[5], Rui Vitorino[6], Pedro Domingues[6], Gilberto Igrejas*[1,2].

[1]Department of Genetics and Biotechnology, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal; [2]Institute for Biotechnology and Bioengineering, Centre of Genetics and Biotechnology, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal; [3]Centre of Studies of Animal and Veterinary Sciences, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal; [4]Veterinary Science Department, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal; [5]Biochemistry and Molecular Biology Area, University of Rioja, Logroño, Spain; [6]Chemistry Department, University of Aveiro, Aveiro, Portugal.

Received: 8 May 2010 **Accepted:** 17 May 2010 **Available Online**: 29 June 2010

ABSTRACT

Acquired resistance to beta-lactams is mainly mediated by extended-spectrum beta-lactamases (ESBLs) that confer bacterial resistance to all beta-lactams except carbapenems and cephamycins, which are inhibited by other beta-lactamase inhibitors such as clavulanic acid. Although ESBLs still constitute the first cause of resistance to beta-lactams among Escherichia coli, other "new beta-lactamases" conferring resistance to carbapenems, such as metallo-beta-lactamases (MBL) and KPC carbapenemases, or to cephamycins, such as CMY enzymes, have more recently emerged and are often associated with ESBLs. In order to identify and characterize the proteome of extended-spectrum β-lactamase (ESBL) type TEM-52 and CMY-2 producing-Escherichia coli strains of human clinical origin a bidimensional electrophoresis (2-DE) technique with an isoelectric focusing followed by a SDS-PAGE, were used. Full proteomic studies were conducted in the same IEF and SDS-PAGE conditions, for two protein samples of E. coli strains with similar antibiotic-resistance profiles recovered from human clinical sources. A total of 64 and 91 spots were recovered and identified in C583 and C580 strains, respectively. Our results will be helpful for further understanding of antibiotic-resistant mechanism.

Keywords: Antibiotic resistance; ESBL; *Escherichia coli*; Proteome; Humans.

1. Introduction

Proteomics is the large-scale study of proteins, particularly their structures and functions [1]. After genomics, proteomics is often considered the next step in the study of biological systems. It is much more complicated than genomics mostly because while an organism's genome is more or less constant, the proteome differs from cell to cell and from time to time. This is because distinct genes are expressed in distinct cell types. This means that even the basic set of proteins which are produced in a cell needs to be determined. In the past this was done by mRNA analysis, but this was found not to correlate with protein content. It is now known that mRNA is not always translated into protein, and the amount of protein produced for a given amount of mRNA depends on the gene it is transcribed from and on the current physiological state of the cell. Proteomics confirms the presence of the protein and provides a direct measure of the quantity present [2].

Bacterial surface proteins are important for the host-pathogen interaction and they are frequently involved in disease pathogenesis. A wide variety of bacterial surface proteins is represented by lipoproteins which are important components of the bacterial transport system; transmembrane structures involved in the import-export of substrates, including sugars, amino acids, oligopeptides, polyamines, various metal ions and minerals. These systems contribute to many bacterial processes, such as acquisition of nutrients, stress responses

***Corresponding author:** Dr. Gilberto Igrejas. Department of Genetics and Biotechnology, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal. Email address: gigrejas@utad.pt

and intercellular signalling, many of which could be vital for bacterial growth and survival within the host [3]. In the last years extended-spectrum β-lactamases (ESBLs) of the CTX-M class are dramatically spreading among human clinical E. coli isolates, mainly in those recovered from the community [4]. Escherichia coli have become one of the most important causes of nosocomial and community acquired infections. Beta-lactams (mainly extended-spectrum cephalosporins and carbapenems) and fluoroquinolones constitute the main therapeutic choices to treat infections caused by these micro-organisms. However, resistance to these compounds has been reported more and more frequently in Europe in the past years [5]. However, the increasing number of Enterobacteriaceae with ESBLs that also contain AmpCs and other new mechanisms of resistance to fluoroquinolones or aminoglycosides indicate that the recent increase of ESBL producers in Europe constitutes a complex multifactorial problem of high public health significance that deserves a deep analysis and the implementation of specific interventions at different levels.

Two-dimensional gel electrophoresis (2-D electrophoresis) is a powerful and widely used method for the analysis of complex protein mixtures extracted from cells, tissues, or other biological samples. It allows a fast overview of changes in cell processes by analysis of the entire protein extracts in any biological and medical research projects. New instrumentation and advanced technologies provide proteomics studies in a wide variety of biological and biomedical questions. Proteomics work is being applied to study antibiotics-resistant strains and human tissues of various brain, lung, and heart diseases [6]. This technique separate proteins in two steps, according to two independent properties: the first-dimension is isoelectric focusing (IEF), which separate proteins according to their isoelectric points (pI); the second-dimension is SDS-polyacrylamide gel electrophoresis (SDS-PAGE), which separates proteins according to their molecular weights (MW). In this way, complex mixtures consisted of thousands of different proteins can be resolved and the relative amount of each protein can be determined. The procedure involves placing the sample in gel with a pH gradient, and applying a potential difference across it. In the electrical field, the protein migrates along the pH gradient, until it carries no overall charge. This location of the protein in the gel constitutes the apparent pI of the protein. There are two alternatives methods to create the pH gradient - carrier ampholites and immobilized pH gradient (IPG) gels. [7]. Mass spectrometry (MS) is a powerful tool in protein analysis. Electrospray and matrix-assisted laser desorption ionization (MALDI) time-of-flight (TOF) technologies can be used to precisely detect small changes in the masses of proteins and peptides. These techniques involve the ionization of molecules into products that can be detected. The mass-to-charge ratio of gas phase ions can then be correlated with the molecular structure of the initial species. Electrospray ionization involves an electric field applied to a solution sprayed from a needle. In MALDI, gas phase ions are generated by desorption ionization of the molecule of interest from a layer of crystals formed from volatile matrix molecules. By using these techniques mutated proteins can be detected rapidly, and the precise site of the mutation can be characterized using tandem MS/MS of peptides of the protein [8]. Antibiotic resistance presents a significant challenge to scientists in the field of infectious diseases. Identification of protein determinants for resistance will not only provide markers for resistance to a particular drug but will also aid in the understanding of the mechanisms of antibiotic function and resistance. Several antibiotics act by targeting protein biosynthesis, interacting with ribosomal structural proteins, rRNAs, and ribosomal-associated proteins [9].

In the current study, a proteomic evaluation of two E. coli isolates (C583 and C580), recovered from clinical human samples and carrying a plasmidic β-lactamase of class AmpC (CMY-2) and a TEM-52 enzyme, respectively , was performed by 2-DE and subsequent protein identification by MALDI-TOF MS.

2. Material and methods

2.1 *Cell culture and purification of E. coli*

ESBL-producing E. coli strains (C583 and C580) of human clinical origin were previously studied and characterized by Vinué et al. [10] and were included in this study. Consequently, proteomic analysis was performed with E. coli strains from human clinical samples.

2.2 *Protein extraction*

Exponentially growing cells (15 mL) were harvested by centrifugation (3 min, 10,000 xg, 4ºC) and re-suspended in PBS (4 mL) at room temperature, followed by a second centrifugation and resuspension with SDS+Tris solution (0.2 mL) [11]. Cell disruption was performed by sonication (3 × 10 s, 4 °C at 100 W); cell debris was removed by centrifugation (14,000 xg, 30 min at 4ºC). The protein concentration was assayed using a 2D Quant kit (GE Healthcare).

2.3 *One-dimensional electrophoresis and staining*

One-dimensional electrophoresis was conducted with SDS-polyacrylamide gels (T=12.52%, C=0.97%) in a HoeferTM SE 600 Ruby® (Amersham Biosciences) unit, following Laemmli [12] with some specific modifications [13]. Gels were stained during 24 hours in Coomassie Brilliant Blue R-250 and washed in water overnight. It was then fixed in trichloroacetic acid (6%) for four hours and in glycerol (5%) for two hours [14].

2.4 *Two-dimensional electrophoresis and proteome analysis*

2-DE was performed according to the principles of O'Farrell (1975) but with IPG (ImmobilineTM pH Gradient) technology [12]. Protein samples of E. coli (GV5) were used in parallel with samples of E. coli C583 and C580. For IEF, precast IPG strips with linear gradient of pH 3-10 were pas-

sively rehydrated overnight (12 to 16 hours) in a reswelling tray with rehydration buffer (8M urea, 1% CHAPS, 0.4% DTT, 0.5% carrier ampholyte IPG buffer pH 3-10) at room temperature. IPG strips were covered with DryStrip Cover Fluid (Plus One, Amersham Biosciences). Lyses buffer [9.5M urea, 1% (w/v) DTT, 2% (w/v) CHAPS, 2% (v/v) carrier ampholytes (pH 3-10) and 10 mM Pefabloc® proteinase inhibitor] was added to the E. coli isolates (1:1). Samples containing a total of 100 µg of protein were loaded into 13 cm IPG strips (pH 3–10 NL, Amersham Biosciences, UK) [14]. The sample solution was then applied to the previously rehydrated IPG strips pH3-10 by cup loading and then proteins were focused sequentially at 500 V for 1 h, gradient at 1000 V for 8 h, gradient at 8000 V for 3 h, and finally 8000 V during 1 h on an EttanTM IPGPhor IITM (Amersham Biosciences, Uppsala, Sweden). Seven IEF replicate runs were performed according to Görg [15] and the GE Healthcare protocol for IPG strips pH 3-10 of 13 cm, in order to obtain the optimized running conditions, resulting in a final 13 hour run. Focused IPG strips were then stored at -80°C in plastic bags. Before running the second dimension, strips were equilibrated twice for 15 minutes in equilibration buffer (6 M urea, 30% (w/v) glycerol, 2% (w/v) SDS in 0.05 M Tris-HCl buffer (pH 8.8)). In the first equilibration 1% DTT was added to the original equilibration buffer and to the second, 4% iodoacetamide. Bromophenol blue was added to both solutions. The equilibrated IPG strips were gently rinsed with SDS electrophoresis buffer, blotted to remove excessive buffer, and then applied to 12.52% polyacrylamide gels in a HoeferTM SE 600 Ruby® (Amersham Biosciences) unit. Some modifications were introduced in the SDS-PAGE technique previously reported by Laemmli (1970), that allowed its resolution to be increased, with proper insertion of the IPG strips in the stacking gel [12,13]. After SDS-PAGE, the 2-DE gels were fixed in 40% methanol / 10% acetic acid for one hour and afterwards stained overnight in Coomassie Brilliant Blue G-250 [16]. Coomassie-stained gels were scanned on a flatbed scanner (Umax PowerLook 1100; Fremont, CA, USA), and the resulting digitized images were analyzed using Image Master 5.0 software (Amersham Biosciences; GE Healthcare).

2.5 *Protein identification by MALDI-TOF/TOF*

To increase experimental efficacy, four separate gels were analyzed originally representing three independent E. coli protein samples that were previously pooled together and compared. Spots of expression in all gels were manually excised from the gels and analyzed using Matrix-Assisted Laser Desorption/Ionization-Time of Flight Mass Spectrometry (MALDI-TOF). The gel pieces were washed three times with 25mM ammonium bicarbonate /50 % ACN (acetonitrile), once with ACN and dried in a SpeedVac (Thermo Savant). 25 mL of 10 mg/mL sequence grade modified porcine trypsin (Promega) in 25mM ammonium bicabornate was added to the dried gel pieces and the samples were incubated overnight at 37°C. Extraction of tryptic peptides was performed by addition of 10% of formic acid (FA)/50% ACN followed by three-fold lyophilisation in a SpeedVac (Thermo Savant). Tryptic peptides were re-suspended in 10 mL of a 50% acetonitrile/0.1% formic acid solution. The samples were mixed (1:1) with a matrix consisting of a saturated solution of α-cyano-4-hydroxycinnamic acid prepared in 50% acetonitrile/ 0.1% formic acid. Aliquots of samples (0.5µL) were spotted onto the MALDI sample target plate.

Peptide mass spectra were obtained on a MALDI-TOF/TOF mass spectrometer (4800 Proteomics Analyzer, Applied Biosystems, Europe) in the positive ion reflector mode. Spectra were obtained in the mass range between 800 and 4500 Da with ca. 1500 laser shots. For each sample spot, a data dependent acquisition method was created to select the six most intense peaks, excluding those from the matrix, trypsin autolysis, or acrylamide peaks, for subsequent MS/MS data acquisition. Mass spectra were internally calibrated with autodigest peaks of trypsin (MH+: 842.5, 2211.42 Da) allowing a mass accuracy of better than 25 ppm.

2.6 *Database search*

Spectra were processed and analyzed by the Global Protein Server Workstation (Applied Biosystems), which uses internal MASCOT software (v 2.1.04, Matrix Science, London, UK) for searching the peptide mass fingerprints and MS/MS data. Swiss-Prot nonredundant protein sequence database (Release 57 of March 2009, 428650 entries) was used for all searches under E. coli. The database search parameters were as follows: carbamidomethylation and propionamide of cysteine (+71Da) as a variable modification as well as oxidation of methionine (+16Da), and the allowance for up to two missed tryptic cleavages. The peptide mass tolerance was 25 ppm and fragment ion mass tolerance was 0.3 Da. Positive identifications were accepted up to 95% of confidence level. Protein identifications were considered as reliable when the MASCOT score was > 70 (MASCOT score was calculated as $-10 \times \log P$, where P is the probability that the observed match is a random event.). This is the lowest score indicated by the program as significant ($P < 0.05$) and indicated by the probability of incorrect protein identification.

3. Results and Discussion

E. coli strain C583 exhibited resistance to ampicillin, amoxicillin/clavulanic acid, cefoxitin, cefotaxime, ceftazidime, naladixic acid, ciprofloxacine, and trimethoprim-sulfametoxazole and harbored the intI1 gene of class 1 integrons. A plasmidic β-lactamase of class AmpC (CMY-2) was detected in this strain. On the other hand, E. coli C580 presented resistance to ampicillin, cefotaxime, ceftazidime, naladixicacid, ciprofloxacin, and trimethoprim-sulfametoxazole but did not contain class 1 integrons. This strain was proven to be a β-lactamase TEM-52 producer [10].

A comparative analysis among the strains was carried out. The protein expressions of E. coli strains were visualized on 2-DE gels (Figure 1). The use of pH 3-10 IPG strips resulted in a well spread display of protein spots which allowed their safe and accurate excision and image identification.

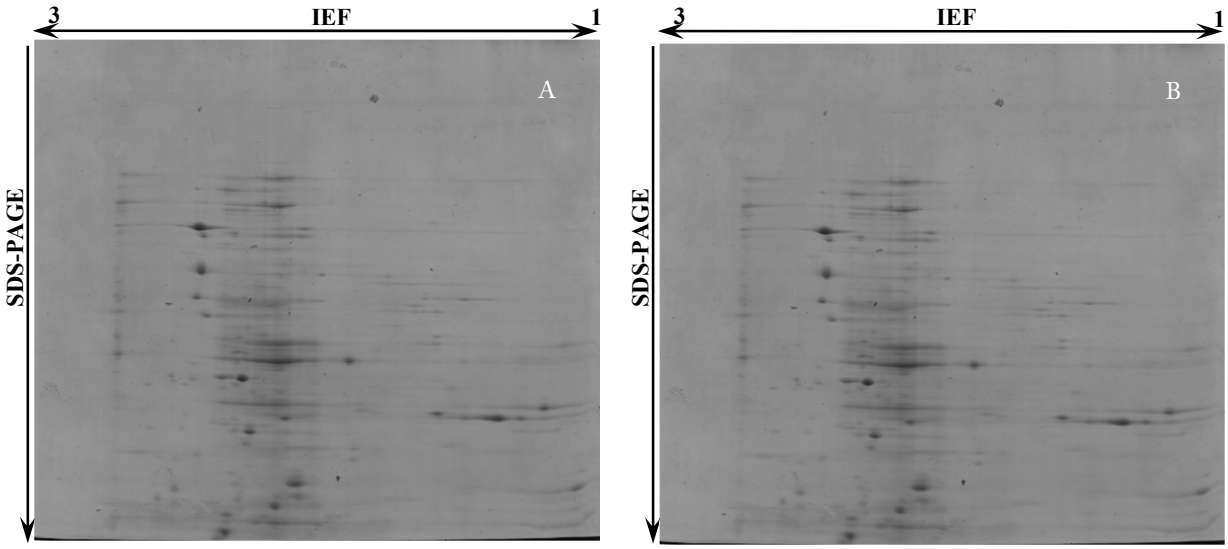


Figure 1. 2-DE gel image of E. coli samples with IPG strips pH3-10. A: C583 B: C580.

Full proteomic studies were conducted in the same IEF and SDS-PAGE conditions, for two protein samples of E. coli strains of human clinical samples C583 and C580. A total of 64 and 91 spots were recovered. All the proteins present in the 2-D gels spots were identified by MALDI-TOF MS and protein bioinformatic databases querying (Tables 1 and 2 of the supplementary material). Among the proteins identified, it is important to emphasise the presence of proteins related to biosynthesis and regulation, glycolysis, stress response, cellular metabolic processes and antibiotic resistance in both

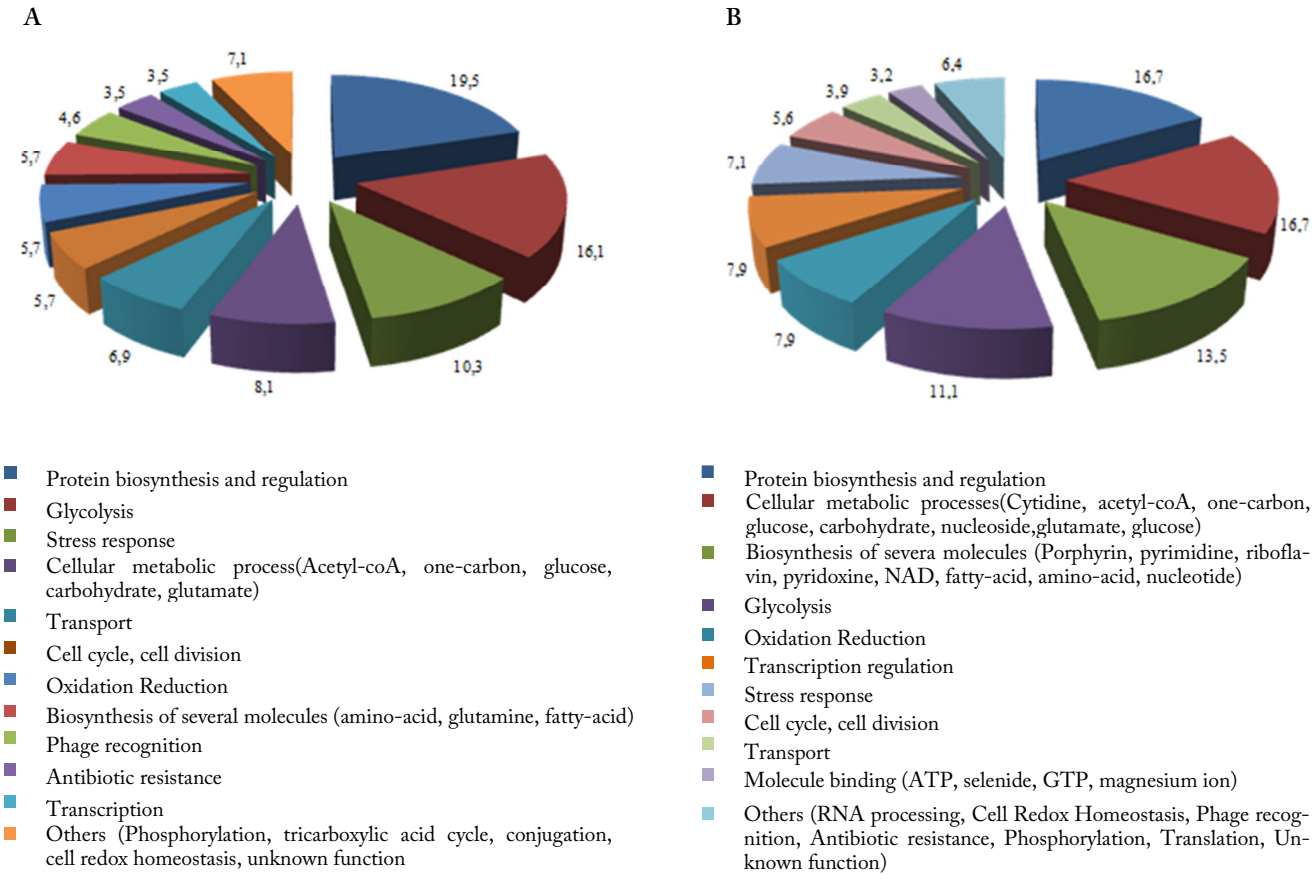


Figure 2. Distribution of the biological processes related to the protein spots found in the 2-DE gels of the *E. coli.* A: C583 and B: C580.

E. coli strains (Figure 2). Proteins related to nine different E. coli strains were found in samples of our C580 and C583 strains; namely proteins associated to three commensal strains (K12, O9:H4 and strain ATCC 8739 / DSM 1576 / Crooks), two uropathogenic (UTI89 and O6:K15:H31 (strain 536)), one avian pathogenic (O1:K1), one enterohemorrhagic (O157:H7), one enterotoxigenic (O139:H28 (strain E24377A)) and one virulent strain (O6) were identified in our strains. The presence of proteins related to one virulent strain O6 and one enterohemorrhagic strain O157:H7 is consistent with the results found in the seagull isolates previously studied in this report. Related to virulent E. coli strain O6 were identified two proteins for E. coli isolate C583 (Trigger factor protein and Chaperone protein ClpB) and three proteins for E. coli isolate C580 (Trigger factor protein, Chaperone protein ClpB and Pyridoxine 5'-phosphate synthase protein). In the case of enterohemorrhagic E. coli strain, two related proteins were found in both E. coli isolate C583 (Putative flavoprotein and Serine hydroxymethyltransferase protein) and other two were found in E. coli C580 (Putative flavoprotein and Curved-DNA binding protein). Also noticed was the detection of proteins Malate dehydrogenase and GrpE protein linked with E. coli enterohemorrhagic strain O157:H7 in our human isolates. A BLAST of a putative wrbA flavoprotein isolated from C580 sequence was released with Translate Reverse tool (Figure 3).

Uropathogenic strains are generally involved in urinary tract infections (UTI) in humans, with UTI89 strain being related to the expression of adhesive organelles (known as pili) that interact with proteins of urinary epithelial cells [17].

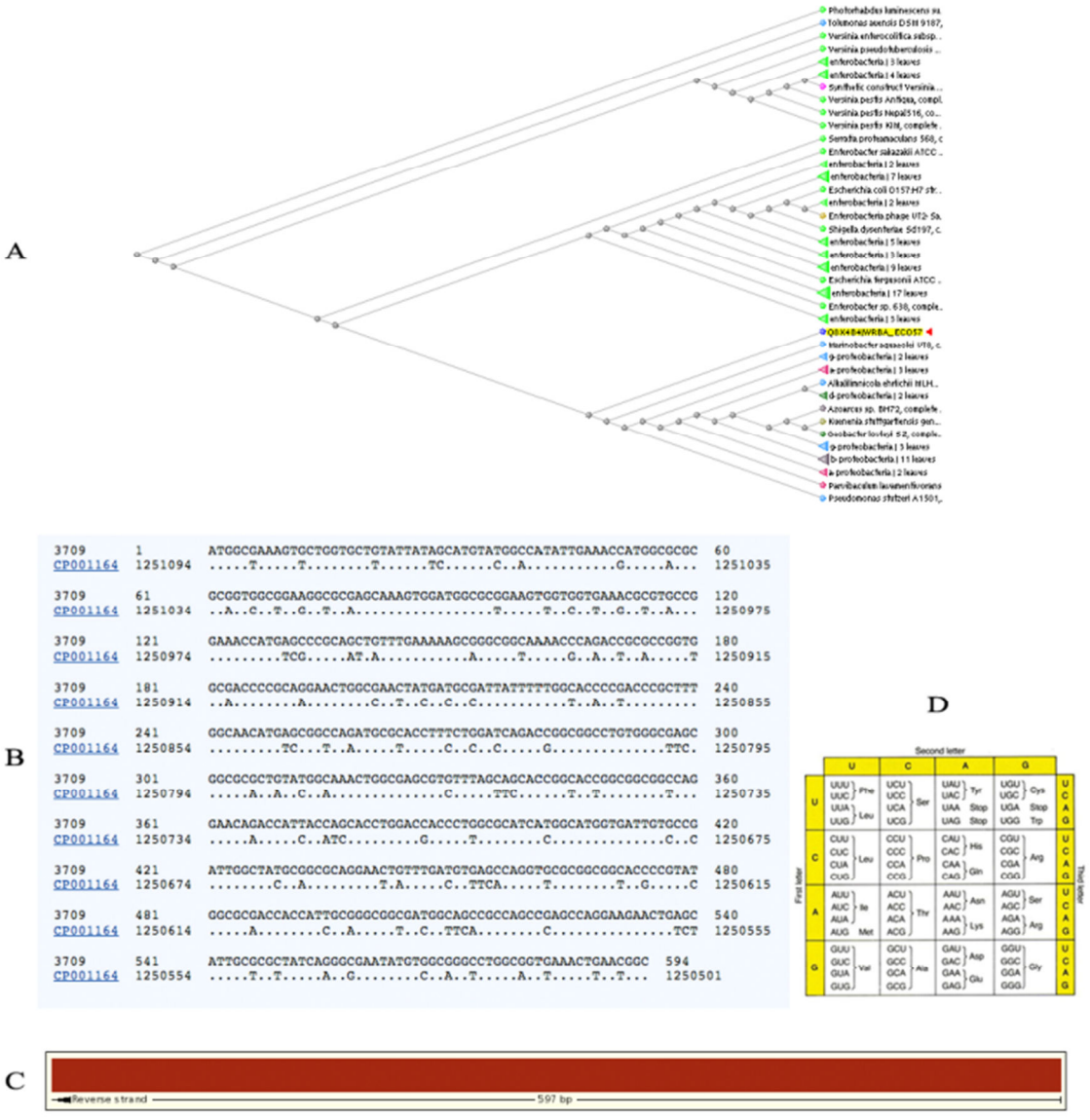


Figure 3. BLAST of a putative *wrbA* flavoprotein isolated from C580 *in silico* sequence from *Translate Reverse* tool (http://www. bioinformatics.org/sms2/rev_trans.html). A: Phylogenetic tree in terms of BLAST with BLASTn alghortim from NCBI; B: Alignment with *in silico* sequence and *wrbA* flavoprotein of *E. coli* O157:H7 EC4115; C: Analysis of a transcript of *wrbA* flavoprotein *E. coli* O157:H7 EC4115 gene with *Emsembl Bacteria* (http://bacteria.ensembl.org/ index.html); D: Genetic code.

In the case of E. coli O6:K15:H31, it is related to pyelonephritis isolates involved in ascending UTIs [18]. Enterotoxigenic strains are an important cause of intestinal infections in children and adults with genes encoding heat-labile and heat-stable properties carried on plasmids. Enterotoxin plasmids are even among the first to be associated with virulence and E. coli O139:H28 (strain E24377A) is among the current examples of this type of E. coli harmful to humans [19]. Avian pathogenic strains such as O1:K1 belong to the same highly pathogenic clonal groups as E. coli strains of the same serotype isolated in several clinical cases of UTIs, neonatal meningitis and septicemia therefore constituting a potential zoonotic risk [20]. Finally, proteins previously characterized as linked with three commensal strains were also found in E. coli isolates from both human faecal samples: K12, O9:H4 and strain ATCC 8739 / DSM 1576 / Crooks. E. coli K12 is a well-known commensal strain with two major porines OmpC and OmpF, corresponding proteins of the external membrane [21]. E. coli strain ATCC 8739 / DSM 1576 / Crooks differs from strain K12 by presenting an insertion element IS1-13 associated to an upstream promoter and to the first 114 bp of OmpC, therefore only expressing OmpF [22]. The E. coli strain O9:H4 was originally isolated from stool from a healthy laboratory scientist who had been using an undomesticated E. coli isolate for several human colonization studies. With no damage done to the hosts, this strain is normally used as a control sample for comparison of pathogenical strains [23]. Such a large number of proteins related to different E. coli strains present in the same faecal sample should cause concern as it represents a major antibiotic resistance at several levels associated with the presence of proteins found in both humans. The relationship between the resistance mechanisms in clinical strains and the understanding of its functioning is of great value for public health worldwide. The full knowledge of how antibiotics resistance evolves and are transmitted between potential hosts of different ecosystems takes on great importance. Proteomic analysis and protein identification can become important and reliable complementary tools to improve our knowledge in this field.

4. Concluding remarks

A well-defined E. coli proteome will have direct applications in biochemical, biological, and biotechnological research fields in the following ways: (i) underpins our understanding not only of the prokaryotic regulatory network but also of complex eukaryotic regulatory networks including stimulon, regulon, and cascade-like networks, (ii) provides invaluable information for designing metabolic engineering strategies to enhance production of various bioproducts, including recombinant proteins, biopolymers, and metabolites, (iii) can be used as a model system to help accelerate the development of advanced high-resolution, high-throughput, and high-sensitivity proteomic technologies [24]. The evaluation of protein profiles in response to various stress mechanisms, such as sensitivity to antibiotics or modifications related to antibiotic resistance, could represent a valid and integrating approach for the development of new therapeutic strategies [25]. In this study, the elaboration of a 2-DE electrophoresis gel of an extended-spectrum β-lactamase E. coli strain with phenotypic and genotypic profiles indicating antimicrobial resistance allowed us to identify and characterize the proteins present. Considering that the worldwide emergence of antibiotic-resistant bacteria poses a serious threat to human health, understand mechanisms of the resistance are extremely important to the control of these bacteria. Our findings will be helpful for further understanding of antibiotic-resistant mechanism(s).

5. Supplementary material

Supplementary material regarding this manuscript is online available in the web page of JIOMICS.

http://www.jiomics.com/index.php/jio/rt/suppFiles/20/0

References

1. S.D. Patterson, R.H. Aebersold, Nat. Genet. 33 (2003) 311–323.
2. E.A. Panisko, T.P. Conrads, M.B. Goshe, T.D. Veenstra, Exp. Hematol. 35 (2002) 97-107.
3. P. Roncada, F. Deriu, A. Gaviraghi, P. A. Martino, L. Bonizzi, Vet. Res. Commun. 33 (Suppl 1) (2009) S157–S160.
4. G.A. Jacoby, L.S. Munoz-Price, New England Journal of Medicine 352 (2005) 380-391.
5. F. Perez, A. Endimiani, K.M. Hujer, R.A. Bonomo, Curr. Opin. Pharmacol. 34(11) (2004) 514-521.
6. B. Wittmann-Liebold, H.-Rüdiger Graack, and T. Pohl. Proteomics 6 (2006) 4688–4703.
7. http://www.tau.ac.il/lifesci/units/proteomics/.
8. S.K. Wilcox, G.S. Cavey, J.D. Pearson. Antimicrob. Agents Chemother. 45(11) (2001) 3046–3055.
9. C.M.T. Spahn, C.D. Prescott. J. Mol. Med. 74 (1996) 423–439.
10. L. Vinué, M. Lantero, Y. Sáenz, S. Somalo,I. de Diego, F. Pérez, F. Ruiz-Larrea, M. Zarazaga, C. Torres, J. Med. Microbiol. 57 (2008) 916-920.
11. J.E. Celis, P. Gromov, Curr. Opin. Biotechnol. 10(1) (1999) 16-21.
12. U.K. Laemmli, Nature 227 (1970) 680-685.
13. G. Igrejas, Genetic, biochemical and technological factors associated to the utilization of common wheat (Triticum aestivum L.), University of Trás-os-Montes and Alto Douro 2000, Ph.D. Thesis.
14. A. Görg, C. Obermaier, G. Boguth, A. Harder, B. Scheibe, R. Wildgruber, W. Weiss, Electrophoresis 21 (2000) 1037-1053.
15. A. Görg, A. Klaus, C. Lück, F. Weiland, W. Weiss, Technische Universität München 2007.
16. A. Görg, W. Weiss, M.J. Dunn, Proteomics 4 (2004) 3665-3685.
17. D.A. Hunstad, S.S. Justice, C.S. Hung, S.R. Lauer, S.J. Hultgren, Infect. and Immun. 73(7) (2005) 3999-4006.
18. A.L. Lloyd, D.A. Rasko, H.L.T. Mobley, J. Microbiol. 189(9) (2007) 3532-3546.
19. M.K. Wolf, Clin. Microb. Rev. 10(4) (1997) 569-584.
20. M. Moulin-Schouleur, M. Répérant, S. Laurent, A. Brée, S. Miqnon-Grasteau, P. Germon, D. Rasschaert, C. Schouler, J. Clin. Microbiol. 45(10) (2007) 3366-3376.
21. K. Hayashi, N. Morooka, Y. Yamamoto, K. Fujita, K. Isono, S. Choi, E. Ohtsubo, T. Baba, B.L. Wanner, H. Mori, T. Horiuchi,

Mol. Syst. Biol. 2 (2006) 1-5.

22. A. Copeland, S. Lucas, A. Lapidus, T. Glavina Del Rio, E. Dalin, H. Tice, D. Bruce, L. Goodwin, S. Pitluck, H. Kiss, T. Brettin, J.C. Detter C. Han, C.R. Kuske, J. Schmutz, F. Larimer, M. Land, L. Hauser, N. Kyrpides, N. Mikhailova, L. Ingram, P. Richardson, Direct submission.
23. D.A. Rasko, M.J. Rosovitz, G.S. Myers, E.F. Mongodin, W.F. Fricke, P. Gajer, J. Crabtree, M. Sebaihia, N.R. Thomson, R. Chaudhuri, I.R. Henderson, V. Sperandio, J. Ravel, J. Bacteriol. 190(20) (2008) 6881-6893.
24. M. Han, S.Y. Lee, Microbiol. Molec. Biol. Rev. 70(2) (2006) 362-439.
25. D. Hiscock, C. Upton, Bioinformatics 16(5) (2000) 484-485.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.21

A new method for measuring functional similarity of microRNAs

Guangchuang Yu[1,a], Chuan-Le Xiao[1,a], Xiaochen Bo[2], Chun-Hua Lu[1,3], Yide Qin[4], Sheng Zhan[1], Qing-Yu He*[1].

[1]Institutes of Life and Health Engineering, Jinan University, 601 Huangpu Ave. W. Guangzhou, 510632, China; [2]Beijing Institute of Radiation Medicine, Beijing, 100850, China; [3]College of Life Science and Technology, Guangxi University, Nanning, 530004, China; [4]Department of Biochemistry and Molecular Biology, Anhui Medical University, Hefei, 230032, China. [a] These authors contributed equally in this work.

Received: 28 July 2010 **Accepted:** 18 August 2010 **Available Online**: 10 September 2010

ABSTRACT

MicroRNAs (miRNAs) are a group of small RNAs with regulatory roles at post-transcriptional level. Although they have been clustered based on their sequence or structure similarities, there is still no effective method to determine their functional similarities due to the lack of miRNAs functional annotation. To address this critical need, we presented here a novel method for systematic study of functional similarities among human miRNAs by using their target genes GO semantic similarities. The functional similarities were validated by comparing with miRNA expression similarities. To extract the highly significant clusters, we used multi-scale bootstrap re-sampling in clustering miRNAs functional similarities. The clustering of human miRNAs based on target genes molecular function annotation led to 44 significant clusters. The clustering results were coherent with biological knowledge. Our analysis suggests that systematic clustering based on target genes GO semantic similarities can aid to reveal the functional diversity of miRNA families. Additionally, this method can be extended to other species and used to predict novel miRNA functions.

Keywords: MicroRNA; Gene Ontology; Semantic Similarity; Clustering.

1. Introduction

MicroRNAs (miRNAs), approximately 22 nucleotides in length, are non-coding RNAs that play crucial roles in post-transcriptional regulation. MiRNAs are evolutionarily conserved, and generally transcribed by RNA polymerase II. MiRNAs perform their functions by RNA induced silencing complex (RISC), leading their target mRNAs to direct destructive cleavage or translational repression. MiRNAs are considered to represent one of the most important components of the cell. They involve in many critical biological processes, including cell development and differentiation [1,2], proliferation [3], apoptosis [4], development [5], immune system regulation [6,2], cancer progression [7], and virus-host interaction [8] and therefore represent potential targets for therapeutic applications. According to miRBase [9], the number of registered miRNA genes continues to grow rapidly. However, hundreds of recently identified miRNAs have unknown functions due to the lack of experimental strategy for systematic identification of their regulating targets.

In order to better understand miRNAs, it is increasingly necessary to measure their functional similarity and thus to infer novel potential functions for miRNAs. Human miRNAs have been grouped into 46 families on the basis of hairpin sequences conservation by Rfam [10], and 60 families according to pre-miRNAs sequence and secondary structures by using FOLDALIGN [11]. Many human clusters containing miRNAs without sequence homology was found [12] which indicating the current strategies for measuring miRNA similarity have some flaws. Since miRNA-mRNA duplex allows mismatch, and target recognition only matches the 6-8nt long seed region, miRNAs with similar sequences and pre-miRNAs with similar structures may have distinctive functions. Therefore, a new method for measuring miRNAs functional similarity is necessitated.

Gene Ontology (GO) is the *de facto* standard for annotation of gene products. The relationship of different genes was

***Corresponding author:** Qing-Yu He, Institutes of Life and Health Engineering, Jinan University, 601 Huangpu Ave. W. Guangzhou, 510632, China. Email address: tqyhe@jnu.edu.cn

organized as a directed acyclic graph (DAG), making it feasible for quantitative semantic comparisons. Measuring gene functional similarity based on GO has been widely used in novel GO annotations [13,14], gene function prediction [15] and similarity measurement [16,17], microarray analysis [18], cluster analysis [19,20], pathway analysis [21,22], and localization prediction [23]. The powerfulness of GO semantic similarity measurement has been verified in terms of the correlations with sequence similarities [24], gene expression profiles [25], and protein-protein interactions [26]. However, it is impractical to measure miRNAs functional similarity directly due to the lack of GO terms annotation of most miRNAs and the lack of functional annotation database [27] since most miRNAs functions remain unknown. Fortunately, most of the genes miRNAs regulated are well annotated, making it possible for inferring functional similarity of miRNAs based on their targets. The functions of miRNAs can be inferred by GO enrichment analysis of their predicted targets were reported [28,29], which also indicating that the functional similarity of miRNAs can be measured by quantitatively calculating the similarity of their target genes.

Here, we proposed a new method for systematic study of miRNAs functional similarity. The fundamental concept of our method is based on the functions of miRNAs targets GO semantic similarities. We validated our method by comparing it with miRNA expression similarity and showed that clustered miRNAs have functional relatedness through co-function of targeting genes. The similarities obtained by our method are consistent with biological knowledge of miRNA functional relationship.

2. Material and methods

2.1 *Human microRNA target prediction*

Experimental identification of target mRNAs is difficult, and TarBase [30] currently lists only 1093 verified target mRNAs for 110 human miRNAs. Due to the lack of experimental targets of miRNAs, we used computational method for miRNA genome-wide target prediction in this study, where target genes were predicted by the algorithm of Probability of Interaction by Target Accessibility (PITA) [31]. PITA uses a thermodynamic model for miRNA-mRNA interaction that was scored by an energy score, ΔΔG, which equals to the difference between the energy expended on opening the target site structure, ΔGopen, and the energy gained by forming the duplex, ΔGduplex. ΔΔG correlates well with the experimentally measured degree of mRNA suppression were reported [31].

Human miRNAs were downloaded from miRBase, version 12 [9] and human genome was downloaded from UCSC, version 18 [32] which corresponds to the human genome build 36.1 assembled by NCBI. We extracted 3' untranslated region (UTR) sequences in a single FASTA format file. For genes missing 3'-UTR annotation, 800bp downstream annotated end of the coding sequence were used as the putative UTR. Since miRNA-target interaction requires unpairing of bases flanking the targets, we used a flank of 3 upstream and 15 downstream nucleotides when performing prediction. To reduce false positive, the prediction results were narrowed down using the criteria of 7-8 bases seed length, no G:U wobble or loops, no mismatch and conservation score of 0.9 or higher. To assign an overall miRNA-target score, we computed the statistical weight to sum all the ΔΔG generated by different sites bounded to the same miRNA as defined in formula 1. Finally we screened out the results by ΔΔG below -10 kcal/mol. ΔΔG is an energetic score, and the lower its value, the stronger of miRNA-target binding. ΔΔG < -10 kcal/mol is expected to be functional in endogenous miRNA expression levels.

$$Target\_Score = -\log(\sum(\exp(-\Delta\Delta G))) \qquad (1)$$

2.2 *Human microRNA functional similarities measurement*

For measuring functional similarities among miRNAs, we used GO annotation of their target genes to define the similarity index.

Here, we developed an R package called *GOSemSim* [33], which implemented five methods proposed by Resnik [34], Lin [35], Jiang [36], Schlicker [37], and Wang [38] respectively, and was extended to support 19 species, including Anopheles, Arabidopsis, Bovine, Canine, Chicken, Chimp, Coelicolor, E coli strain K12 and Sakai, Fly, Human, Malaria, Mouse, Pig, Rhesus, Rat, Worm, Xenopus, Yeast, and Zebrafish. The program is freely distributed under GPL2 and can directly be installed from Bioconductor. The manual and source code are available at http://bioconductor.org/packages/release/bioc/html/GOSemSim.html. Since GOSemSim package only supports Entrez Gene identifier for measuring similarities among human genes, we used Bioconductor package biomaRt [39] to query BioMart [40] databases for mapping RefSeq identifiers to Entrez Gene identifiers. Molecular function ontology was used to annotate target genes, and Wang method was used to calculate similarity. Wang method was based on GO graph structure, and outperformed other algorithms based on information content and thus being more consistent with human perspectives [38]. In order to give a single distance between two miRNAs, we combined similarity scores of multiple target genes as defined in formula 2. Similarities between two gene sets associated with two miRNAs form a matrix. The similarity of these two miRNAs is the average of maximum row scores and column scores. We used this strategy and finally obtained pairwise semantic similarities among human miRNAs.

$$Sim = \frac{\sum_{1\le i\le m} \max(Score_{[i,]}) + \sum_{1\le j\le n} \max(Score_{[,j]})}{m+n} \qquad (2)$$

2.3 *Hierarchical clustering uncertainty assessment*

The resulting similarity scores were then clustered by R package *Pvclust* [41]. When performing clustering, *pvclust* used multi-scale bootstrap re-sampling to estimate the uncertainty of cluster analysis which has been popular in phylogenetic analysis. The agglomerative method, average linkage, was used, and 10,000 bootstrap replications were run, with relative sample size from 0.5 to 1.4, incrementing in steps of 0.1, for testing *p*-values. For a cluster with approximately unbiased (AU) *p*-value > 0.95, the hypothesis with "the cluster does not exist" is rejected with significance level 0.05. Roughly speaking, these clusters not only "seem to exist" attributed to sampling error, but be stably observed when we increase the number of observation [41]. All clusters in this study were extracted with *p*-value > 0.95.

3. Results and Discussion

We used PITA algorithm to predict human miRNA targets, followed by a strict criteria to control the FDR, and then measured miRNA functional similarity by *GOSemSim* package. As a result, we obtained the pairwise functional similarity of 533 miRNAs which was provided as Supplemental File 1. The pairwise functional similarity of human miRNA repertoire was shown in Figure 1. The full size of Figure 1 was provided as Supplemental File 2.

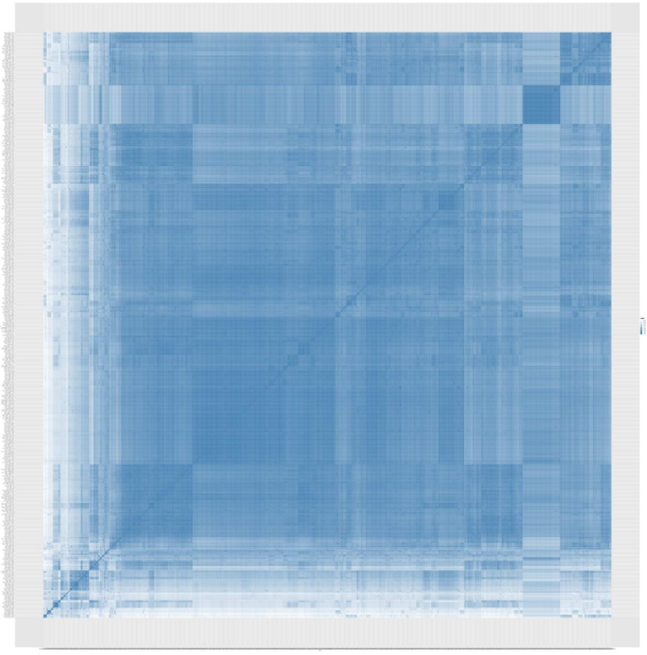


Figure 1. Human miRNAs functional similarities.

MiRNAs with high functional similarity may tend to have similar expression profiles. Taking miRNA functional similarity in pairs against expression similarity should show a positive relationship. For further evaluating the quality of our result, we study the relationship of miRNAs functional similarity and expression similarity. We used Pearson's correlation coefficients to calculate miRNA expression similarity by using miRNA expression profiles of 345 miRNAs in 40 normal human tissues obtained from Liang's investigation [42]. We classified miRNA pairs into separate groups according to functional similarity values by a step of 0.1, and calculated the average of functional similarity and expression similarity of each group, and then measured the correlation of functional similarity with expression similarity (see Supplemental File 4). As expected, functional similarity obtained by our method showed positive correlation with expression similarity (r = 0.6055), in which the functions of miRNAs can partially be explained by their expression level (r^2 = 0.3666).

The pairwise miRNAs functional similarity matrix was then clustered by *Pvclust* package. We obtained 44 *Pvclust* clusters with AU *p*-value > 0.95, containing 401 miRNAs. The result of cluster analysis with high *p*-value highlighted was provided as Supplemental File 3.

Many miRNAs were reported to be associated with diseases. It has been reported that miRNAs implicated in similar disease often have similar functions [27]. In our results, miRNAs associated to similar diseases were more likely to be grouped in the same cluster. For instance, in cluster 23, hsa-miR-215 [43], hsa-miR-221 [44], hsa-miR-194 [45], hsa-miR-193b [46], and hsa-miR-429 [47] were all reported to be associated with adenocarcinoma. In cluster 44, hsa-miR-453 and hsa-miR-219 [48], hsa-miR-93 [49] were reported to relate with breast cancer; while in cluster 29, hsa-miR-30a, hsa-miR-150, hsa-miR-223, and hsa-miR-600 were implicated to relate with Lupus Vulgaris [50].

MiRNAs conserved in evolution may regulate the cardinal biological process cooperatively. We identified that these miRNAs are tend to cluster together. For example, in cluster 24, hsa-miR-300, hsa-miR-495, hsa-miR-154, and hsa-miR-496 were reported to be conserved in genomes of human, chimp, mouse, rat, dog, and cow [51]. In cluster 33, hsa-miR-410, hsa-miR-377, hsa-miR-668, and hsa-miR-381 are conserved in many mammalian genomes and believed to act cooperatively [51]. Transposable elements (TEs) contribute to the evolution dynamics of miRNAs. We found that hsa-miR-325 and hsa-miR-545 derived from TE L2 [52] were grouped to cluster 30.

Many mammalian viruses have been shown to modulate the expression of host cellular miRNAs [53]. MiRNA expression profiles altered by viruses form a novel regulatory layer, and these miRNAs can be grouped to partially reveal the cross-talk between host and virus. In our clustering result, hsa-miR-181a and hsa-miR-15a that were altered in stable hepatitis B virus expressing cell line [54] were identified in cluster 33. Hsa-miR-24 and hsa-miR-638 that were found to have expression changes during in vitro acute hepatitis C virus infection [55] were grouped in cluster 25.

The clustering results obtained by our method were consistent with many other investigations, suggesting that our method is reliable to calculate functional similarities and sensible to cluster miRNAs. The clustering results are useful to reveal functional diversity of miRNA families.

As described previously, the target genes of miRNA were predicted by PITA algorithm, and our current results were consistent with biological data. Therefore, we can predict novel miRNA functions by miRNA pairs with high similarities or by the GO enrichment analysis of the corresponding miRNA cluster. For instance, we predicted novel miRNA functions with similarities above 0.8. Many of them were supported by newly published literature. For example, hsa-miR-107 and hsa-miR-103 regulate lipid metabolism [56], hsa-miR-449 and hsa-miR-203 regulate pRb-E2F1 activity [57, 58], hsa-miR-200b and hsa-miR-429 regulate EGF-driven invasion [59], etc. We also used the cluster information to globally assign the predicted functions to novel miRNAs. GO enrichment analysis were performed in all miRNA clusters, and all the enriched GO terms represent each miRNA cluster with their corresponding *p*-values and other information were provided in Supplemental File 5. Here, we took cluster 23 as an example. After the hypergeometric test, we select the over-represented GO terms of target genes by p-value < 0.001. Consequently, we can identify miRNAs grouped in cluster 23 to have functions of repressing binding, especially the protein binding, metal ion binding and cation binding. Overall, the method presented in this study can be used to predict potential functions of newly identify miRNA and to discover potential miRNAs involved in important pathways.

4. Concluding remarks

In summary, we proposed a novel method for inferring the functional similarities of human miRNAs by integrating the information provided by miRNA target prediction algorithm PITA with Gene Ontology annotation data. The significant miRNA families were also analyzed. This method can be extended to measure miRNA functional similarities of other species. The current method relies on the prediction of target sites that may contain false positives as well as false negatives and thus may bring bias to some extent. However, target prediction algorithms are necessary when predicting functions of newly identify miRNA. In the future, a more robust and reliable miRNA clustering may be obtained when comprehensive experimental miRNA-targets data are available. We believe that with the rapid increase of experimental miRNA-mRNA deposited in TarBase and the improvement of target prediction algorithms, our method will provide high quality miRNA similarity measurement with high sensitivity and specificity.

5. Supplementary material

Supplementary material regarding this manuscript is online available in the web page of JIOMICS.

http://www.jiomics.com/index.php/jio/rt/suppFiles/21/0

Acknowledgements

This work was partially supported by the 2007 Chang-Jiang Scholars Program, Jinan University "211" Projects (Biotechnology & Bioengineering Medicine and Biomaterial & Tissue Engineering).

References

1. Liu, S., Fu, R., Yu, H., Li, K., et al., MicroRNAs Regulation Modulated Self-Renewal and Lineage Differentiation of Stem Cells. Cell Transplantation 2009, 18, 1039-45.
2. Liang, T., Qin, C., The emerging role of microRNAs in immune cell development and differentiation. APMIS: Acta Pathologica, Microbiologica, Et Immunologica Scandinavica 2009, 117, 635-643.
3. Manni, I., Artuso, S., Careccia, S., Rizzo, M. G., et al., The microRNA miR-92 increases proliferation of myeloid cells and by targeting p63 modulates the abundance of its isoforms. The FASEB Journal 2009, 23, 3957-3966.
4. Wang, Y., Lee, C. G. L., MicroRNA and cancer--focus on apoptosis. Journal of Cellular and Molecular Medicine 2009, 13, 12-23.
5. Wienholds, E., Koudijs, M. J., van Eeden, F. J. M., Cuppen, E., Plasterk, R. H. A., The microRNA-producing enzyme Dicer1 is essential for zebrafish development. Nature Genetics 2003, 35, 217-218.
6. Lu, L., Liston, A., MicroRNA in the immune system, microRNA as an immune system. Immunology 2009, 127, 291-298.
7. Cho, W., OncomiRs: the discovery and progress of microRNAs in cancers. Molecular Cancer 2007, 6, 60.
8. Mcclure, L., Kaposi's Sarcoma Herpes Virus Taps into a Host MicroRNA Regulatory Network. Cell Host & Microbe 2008, 3, 1-3.
9. Griffiths-Jones, S., Saini, H. K., van Dongen, S., Enright, A. J., miRBase: tools for microRNA genomics. Nucl. Acids Res. 2008, 36, D154-158.
10. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S. R., Rfam: an RNA family database. Nucleic Acids Research 2003, 31, 439-441.
11. Kaczkowski, B., Torarinsson, E., Reiche, K., Havgaard, J. H., et al., Structural profiles of human miRNA families from pairwise clustering. Bioinformatics (Oxford, England) 2009, 25, 291-294.
12. Yu, J., Wang, F., Yang, G. H., Wang, F. L., et al., Human microRNA clusters: genomic organization and expression profile in leukemia cell lines. Biochem Biophys Res Commun 2006, 349, 59-68.
13. Fontana, P., Cestaro, A., Velasco, R., Formentin, E., Toppo, S., Rapid Annotation of Anonymous Sequences from Genome Projects Using Semantic Similarities and a Weighting Scheme in Gene Ontology. PLoS ONE 2009, 4, e4619.
14. Done, B., Khatri, P., Done, A., Drăghici, S., Predicting novel human gene ontology annotations using semantic analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM 2010, 7, 91-99.
15. Tao, Y., Sam, L., Li, J., Friedman, C., Lussier, Y. A., Information theory applied to the sparse gene ontology annotation network to predict novel gene function. Bioinformatics 2007, 23, i529-538.
16. Mistry, M., Pavlidis, P., Gene Ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics 2008, 9, 327-327.
17. Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E., et al., Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 2008, 9, S4-S4.
18. Cho, Y., Zhang, A., Xu, X., Semantic similarity based feature extraction from microarray expression data. International Journal of Data Mining and Bioinformatics 2009, 3, 333-345.

19. Bolshakova, N., Azuaje, F., Cunningham, P., A knowledge-driven approach to cluster validity assessment. Bioinformatics 2005, 21, 2546-2547.
20. Wolting, C., McGlade, C. J., Tritchler, D., Cluster analysis of protein array results via similarity of Gene Ontology annotation. BMC Bioinformatics 2006, 7, 338-338.
21. Guo, X., Shriver, C. D., Hu, H., Liebman, M. N., Analysis of Metabolic and Regulatory Pathways through Gene Ontology-Derived Semantic Similarity Measures. 2005, 2005, 972-972.
22. Gamalielsson, J., Olsson, B., Gene Ontology-based semantic alignment of biological pathways by evolutionary search. Journal of Bioinformatics and Computational Biology 2008, 6, 825-842.
23. Lei, Z., Dai, Y., Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. BMC Bioinformatics 2006, 7, 491-491.
24. Lord, P. W., Stevens, R. D., Brass, A., Goble, C. A., Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics (Oxford, England) 2003, 19, 1275-1283.
25. Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., et al., Correlation between gene expression and GO semantic similarity. IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM 2005, 2, 330-338.
26. Xu, T., Du, L., Zhou, Y., Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. BMC Bioinformatics 2008, 9, 472-472.
27. Wang, D., Wang, J., Lu, M., Song, F., Cui, Q., Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics 2010, 26, 1644-50.
28. Ren, J., Jin, P., Wang, E., Marincola, F., Stroncek, D., MicroRNA and gene expression patterns in the differentiation of human embryonic stem cells. Journal of Translational Medicine 2009, 7, 20.
29. Roubelakis, M., Zotos, P., Papachristoudis, G., Michalopoulos, I., et al., Human microRNA target analysis and gene ontology clustering by GOmir, a novel stand-alone application. BMC Bioinformatics 2009, 10, S20.
30. Papadopoulos, G. L., Reczko, M., Simossis, V. A., Sethupathy, P., Hatzigeorgiou, A. G., The database of experimentally supported targets: a functional update of TarBase. Nucl. Acids Res. 2009, 37, D155-158.
31. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., Segal, E., The role of site accessibility in microRNA target recognition. Nat Genet 2007, 39, 1278-1284.
32. Pheasant, M., Meyer, L., Hsu, F., Hinrichs, A. S., et al., The UCSC Genome Browser Database: update 2009. Nucleic Acids Research 2009, 37, D755-761.
33. Yu, G., Li, F., Qin, Y., Bo, X., et al., GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 2010, 26, 976-978.
34. Resnik, P., Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artifical Intelligence Research 1998, 11, 130, 95.
35. Lin, D., , Proceedings of the Fifteenth International Conference on Machine Learning, In Proceedings of the 15th International Conference on Machine Learning 1998, pp. 296-304.
36. Jiang, J., Conrath, D., , International Conference Research on Computational Linguistics (ROCLING X) 1997, p. 9008.
37. Schlicker, A., Domingues, F. S., Rahnenführer, J., Lengauer, T., A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics 2006, 7, 302.
38. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., Chen, C., A new method to measure the semantic similarity of GO terms. Bioinformatics (Oxford, England) 2007, 23, 1274-1281.
39. Durinck, S., Spellman, P. T., Birney, E., Huber, W., Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols 2009, 4, 1184-1191.
40. Haider, S., Ballester, B., Smedley, D., Zhang, J., et al., BioMart Central Portal--unified access to biological data. Nucleic Acids Research 2009, 37, W23-27.
41. Suzuki, R., Shimodaira, H., Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 2006, 22, 1542, 1540.
42. Liang, Y., Ridzon, D., Wong, L., Chen, C., Characterization of microRNA expression profiles in normal human tissues. BMC Genomics 2007, 8, 166.
43. Wijnhoven, B. P. L., Hussey, D. J., Watson, D. I., Tsykin, A., et al., MicroRNA profiling of Barrett's oesophagus and oesophageal adenocarcinoma. British Journal of Surgery 2010, 97, 853-861.
44. Park, J. K., Lee, E. J., Esau, C., Schmittgen, T. D., Antisense inhibition of microRNA-21 or -221 arrests cell cycle, induces apoptosis, and sensitizes the effects of gemcitabine in pancreatic adenocarcinoma. Pancreas 2009, 38, e190-e199.
45. Chung, T. K., Cheung, T., Huen, N., Wong, K. W., et al., Dysregulated microRNAs and their predicted targets associated with endometrioid endometrial adenocarcinoma in Hong Kong women. International Journal of Cancer 2009, 124, 1358-1365.
46. Wu, W., Lin, Z., Zhuang, Z., Liang, X., Expression profile of mammalian microRNAs in endometrioid adenocarcinoma. Eur J Cancer Prev 2009, 18, 50-55.
47. Mees, S. T., Mardin, W. A., Wendel, C., Baeumer, N., et al., EP300 - A miRNA-regulated metastasis suppressor gene in ductal adenocarcinomas of the pancreas. International Journal of Cancer 2010, 126, 114-124.
48. Tchatchou, S., Jung, A., Hemminki, K., Sutter, C., et al., A variant affecting a putative miRNA target site in estrogen receptor (ESR) 1 is associated with breast cancer risk in premenopausal women. Carcinogenesis 2009, 30, 59-64.
49. Zhang, L., Huang, J., Yang, N., Greshock, J., et al., microRNAs exhibit high frequency genomic alterations in human cancer. Proceedings of the National Academy of Sciences 2006, 103, 9136-9141.
50. Dai, Y., Sui, W., Lan, H., Yan, Q., et al., Comprehensive analysis of microRNA expression patterns in renal biopsies of lupus nephritis patients. Rheumatol Int 2009, 29, 749-754.
51. Glazov, E. A., McWilliam, S., Barris, W. C., Dalrymple, B. P., Origin, Evolution, and Biological Role of miRNA Cluster in DLK-DIO3 Genomic Region in Placental Mammals. Mol Biol Evol 2008, 25, 939-948.
52. Piriyapongsa, J., Marino-Ramirez, L., Jordan, I. K., Origin and Evolution of Human microRNAs From Transposable Elements. Genetics 2007, 176, 1323-1337.
53. Scaria, V., Hariharan, M., Pillai, B., Maiti, S., Brahmachari, S. K., Host-virus genome interactions: macro roles for microRNAs. Cellular Microbiology 2007, 9, 2784-2794.
54. Liu, Y., Zhao, J. J., Wang, C. M., Li, M. Y., et al., Altered expression profiles of microRNAs in a stable hepatitis B virus-expressing cell line. Chin Med J (Engl) 2009, 122, 10-14.
55. Liu, X., Wang, T., Wakita, T., Yang, W., Systematic identification of microRNA and messenger RNA profiles in hepatitis C virus-infected human hepatoma cells. Virology 2010, 398, 57-67.
56. Wilfred, B. R., Wang, W., Nelson, P. T., Energizing miRNA research: a review of the role of miRNAs in lipid metabolism, with a prediction that miR-103/107 regulates human metabolic pathways. Molecular genetics and metabolism 2007, 91, 209-

217.

57. Yang, X., Feng, M., Jiang, X., Wu, Z., et al., miR-449a and miR-449b are direct transcriptional targets of E2F1 and negatively regulate pRb–E2F1 activity through a feedback loop by targeting CDK6 and CDC25A. Genes & Development 2009, 23, 2388-2393.
58. Shi, Y., Zhu, J., Salomoni, P., Tucci, P., Lu, X., A Sino-British frontier workshop of cancer biology. Cell Death Differ 2008, 16, 648-650.
59. Uhlmann, S., Zhang, J. D., Schwäger, A., Mannsperger, H., et al., miR-200bc/429 cluster targets PLCgamma1 and differentially regulates proliferation and EGF-driven invasion than miR-200a/141 in breast cancer. Oncogene 2010, 29, 4297-4306.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.27

Cloud point method applied for the Apolipoprotein A-I extraction from human plasma and its identification by tandem mass spectrometry

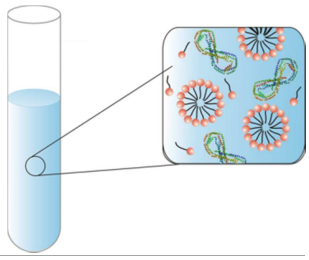
Aline Soriano Lopesa[2,3], Marcelo Anselmo Oseas da Silva[1,2], Fábio Cesar Gozzo[2,4], Marco Aurélio Zezzi Arruda*[1,2].

[1]Group of Spectrometry, Sample Preparation and Mechanization-GEPAM, Department of Analytical Chemistry, Institute of Chemistry, P.O. Box 6154, University of Campinas – UNICAMP, 13083-970, Campinas, SP, Brazil; [2]National Institute of Science and Technology for Bioanalyitics, Institute of Chemistry, University of Campinas – UNICAMP, PO Box 6154, 13083-970, Campinas, SP, Brazil; [3]Butantan Institute – Laboratório Especial de Toxinologia Aplicada – CAT/CEPID, 05503-000, São Paulo, SP, Brazil; [4]Department of Organic Chemistry, Institute of Chemistry, P.O. Box 6154, University of Campinas – UNICAMP, 13083-970, Campinas, SP, Brazil.

Received: 13 July 2010 **Accepted:** 21 August 2010 **Available Online**: 10 September 2010

ABSTRACT

This work describes the extraction of the Apolipoprotein A-I (ApoA-I) from human plasma using the cloud point extraction (CPE). The CPE was carried out with a nonionic surfactant (5% w/v Triton® X-114), and the presence of a salting-out effect (10% w/v NaCl) promoted biocompatible separation conditions at room temperature and pH 6.8. The ApoA-I present in the surfactant-rich phase was identified by tandem mass spectrometry after two-dimensional gel electrophoresis.



Keywords: Protein extraction; Apolipoprotein A-I; Cloud point; Gel electrophoresis; Mass spectrometry.

1. Introduction

Besides other applications [1,2], surfactants are also employed for the extraction of different biomolecules presenting hydrophobic characteristics [3,4]. They allow selective and efficient separation of proteins, especially membrane proteins, which cannot be easily extracted from sample matrix using milder treatments, exploiting the cloud point (CP) phenomenon [5].

CP phenomenon occurs due to the reduction of the surfactant monomers solubility, which is caused by the increase of ionic strength and/or temperature of the solution [6,7,8], resulting in the aggregation of micelles and formation of two apparently immiscible phases, one presenting high surfactant concentration, termed surfactant-rich phase and other with low surfactant concentration, called surfactant-poor phase [9,10]. Due to differences in polarity, hydrophobic proteins are expected to be present in surfactant-rich phase and the hydrophilic ones in the surfactant-poor phase, if dimensional aspects and the concentration of the protein are not taken into account [3,11].

The most common adopted strategy to promote phase separation when extraction of proteins is desired consists in the adjustment of the ionic strength of the medium by adding kosmotropic ions to the solution, exploiting the salting-out effect. This procedure minimizes protein denaturation allowing phase separation at low temperature [12,13]. The use of nonionic surfactants, which separate phases at low temperatures [14], such as some surfactants from Triton® series, can be another alternative for this purpose [15-17].

Different strategies are recently described in the literature demonstrating the feasibility of cloud point extraction (CPE) for the separation of different classes of proteins at laboratorial or industrial scale [18-21]. The application of nonionic molecules consists is the most widespread alternative and can include separations based on the polarity [9], excluded vol-

***Corresponding author:** Marco Aurélio Zezzi Arruda ; Group of Spectrometry, Sample Preparation and Mechanization-GEPAM, Department of Analytical Chemistry, Institute of Chemistry, P.O. Box 6154, University of Campinas – UNICAMP, 13083-970, Campinas, SP, Brazil.

umes [22] or the charge of the proteins [23]. However, few applications are found using complex matrices, such as milk derivates [9,24] and human blood plasma [25].

Particularizing the discussion about blood plasma, this matrix presents proteins with different concentrations ranges and physical characteristics. In this way, the selective removal of specific proteins or classes of proteins consists in a fundamental step in sample preparation procedures applied for several studies, such as proteomics [26,27].The application of an optimized CPE procedure can be an interesting alternative for amphiphatic blood proteins isolation due to the interaction of these macromolecules with the hydrophobic interior of micelles using a simple analytical procedure previously optimized. As a model of amphiphatic blood protein, one can mention Apolipoprotein A-I (ApoA-I) which is found in blood as the major component of high-density lipoproteins (HDL) of plasma [28] being its concentration around 100-150 mg dL-1 [29]. Besides acting in lipid binding and formation of HDL, ApoA-I contributes to cellular cholesterol efflux and the activation of lecithin: cholesterol acyltransferase (LCAT) [23].

In this way, the main goal of this work was to develop a simple, efficient and inexpensive CPE method for ApoA-I extraction from human plasma including the further protein characterization using mass spectrometry to point out the potentialities of CPE for purification of amphiphatic proteins from a complex matrix.

2. Material and methods

2.1 *Reagents and samples.*

The nonionic surfactant polyethylene glycol tert-octylphenyl ether (Triton® X-114) was obtained from Sigma-Aldrich (Steinheim, Germany). NaCl, KH2PO4/NaOH buffer and acetone were of analytical grade (Merck, Darmstadt, Germany or J.T. Baker, Phillispsburg, USA). The reagents for electrophoresis were from Amersham Biosciences (Uppsala, Sweden), and the mass spectrometry grade trypsin was from Promega (Madison, USA). All solutions were prepared with deionized water (≥ 18.2 MΩ cm) using a Milli-Q water purification system (Millipore, Bedford, USA).

Human plasma was supplied by the Clinical Hospital of Unicamp, which promoted the control of the absence of disease. The bags of plasma were received in our laboratory with a seal of approval. The plasma was sonicated for 15 min and aliquots of 15 mL were frozen at –18 ºC in separated flasks. During the experiments, defrosted aliquots were never refrozen to avoid protein precipitation.

2.2 *Phase separation procedure and surfactant removal.*

Sodium chloride (0.83 g) was dissolved using a vortex mixer in glass tubes containing 8 mL of 5 % (w/v) Triton® X-114 and 200 µL of 0.1 mol L^{-1} KH_2PO_4 / NaOH at pH 6.8. A volume of 100 µL of plasma was then added, the mixture was homogenized again, and the glass tubes were centrifuged at 1780 g for 10 min to accelerate phase separation. The upper and the lower phases were called surfactant-rich and poor phases, respectively. The temperature was ca. 25 ºC for all experiments.

Surfactant removal for total protein quantification according to the Bradford´s method [30] was carried out as follows: for quantification of total proteins present in the surfactant-rich phase, 1 mL of the phase was added to 8 mL of ice-cold acetone in an ice bath, and incubated on ice for 1 h. The supernatant was removed, the pellet dried at room temperature and dissolved into 1 mL of KH_2PO_4 / NaOH buffer at pH 6.8. Proteins contained in the surfactant-poor phase were directly quantified without removing the surfactant, since that its concentration was considered negligible [5].

After protein quantification, partition coefficients (K_p), defined as the ratio between protein concentration in surfactant-rich and -poor phases, respectively, were determined for each studied system to obtain favorable ApoA-I extraction conditions.

2.3 *Gel electrophoresis separation.*

Sodium dodecyl sulphate gel electrophoresis (SDS-PAGE) was used to evaluate protein fractions present in the surfactant-rich and -poor phases. For protein precipitation, 8 mL of ice-cold acetone was added to 1 mL of the surfactant-rich phase in an ice bath. This mixture was then incubated for 1 h. Then, the supernatant was removed, the precipitate dried at room temperature and dissolved into 1 mL of 0.05 mol L^{-1} Tris-HCl buffer containing 13.6 % (w/v) glycerol, 2.7 % (w/v) SDS, and 5.4 % (v/v) β-mercaptoethanol at pH 6.8. The surfactant-poor phase (500 µL) was directly diluted with 500 µL of the same buffer solution. SDS-PAGE separation employed a lab cast 10 % (w/v) polyacrylamide gel of 1.5 mm (Amersham Biosciences, Uppsala, Sweden). A constant voltage of 90 V and 25 mA were applied overnight for protein separation. Protein masses of 5.2 and 10.5 µg were applied to each lane, considering the surfactant-rich and -poor phases, respectively. The protein marker (#SM0431, MBI Fermentas, Hanover, USA) was used for molar mass estimation.

The presence of ApoA-I in the surfactant-rich phase was also checked by two-dimensional gel electrophoresis (2D-PAGE). For this task, 20 mL of this phase was precipitated with acetone following the precipitation procedure already described. The precipitate was dissolved in a mixture containing 7 mol L^{-1} urea and 2 mol L^{-1} thiourea, and immediately desalted using a PD-10 column (Amersham Biosciences, Uppsala, Sweden) containing SephadexTM G-25 for removing the excess of phosphate, chloride and sodium. Finally, the desalted solution was lyophilized and dissolved using a buffer containing 7 mol L^{-1} urea, 2 mol L^{-1} thiourea, 2 % (w/v) CHAPS, 0.002 % (w/v) bromophenol blue, 0.5 % (v/v) ampholytes (pH 3-10) and 1 % (w/v) DTT. A 250-µL buffered sample was kept overnight in contact with a 13-cm IPG strip (ImmobilineTM DryStrip, Amersham Biosciences, Uppsala, Sweden, pH 3-10). An isoelectric focusing procedure was carried out in a Multiphor II system (GE Healthcare) totalizing 16000 V h. Then, proteins were reduced and alkylated

according to Garcia et al. [31] followed by separation in the second dimension using the same procedure already described for SDS-PAGE.

The gels obtained were stained with colloidal CBB G-250 [32] and scanned using an ImageScanner II (Amersham Biosciences, Uppsala, Sweden) with the densitometer operating at 300 dpi. The software Gel-Pro Analyzer 3.1 (Media Cybernetics, Maryland 20910, USA) was used for analyzing the bands in the SDS-PAGE, and the ImageMaster 2D Platinum 6.0 software (GeneBio, Geneva, Switzerland) was employed for data treatment of the separation by 2D-PAGE.

2.4 *Tandem mass spectrometry analysis.*

Spots manually removed from the gel were digested using a micro SPE plate containing a peptide affinity resin – Montage® In-Gel digestZP kit (Millipore, MA, USA), according to the manufacturer's recommendations.

The digested proteins were analyzed using the dried droplet method [33]. For this task, 1 μL of each sample was mixed with 1 μL of MALDI matrix (1 % w/v α-cyano-4-hydroxycinnamic acid dissolved into 1:1 v/v acetonitrile: water solution, containing 0.1 % v/v TFA). The mixture was spotted to a MALDI plate and dried at room temperature until complete solvent evaporation.

Mass spectra were acquired in a MALDI-Q-TOF Premier mass spectrometer (Waters – Micromass, Manchester, UK), and obtained in the positive mode using a Nd:YAG laser (337 nm) MALDI source. LockMass correction was performed with PEG 800 in the lock mass spot. Argon was used as collision gas and a typical collision energy (34 - 161 eV) was employed. The instrument was controlled by MassLynx 4.1 software.

Peptide mass data were analyzed for corresponding protein matching in the MSDB database with oxidation of methionine as variable modification, carbamidomethylation of cysteine as fixed modification, ± 0.1 Da peptide and fragment mass tolerance, and a maximum of one missed cleavage site setting on the MASCOT Server database search engine (Matrix Science, London, UK) [34]. The significance threshold was set at $P < 0.05$, corresponding, in this case, to a minimum MASCOT score of 47. Additional confirmation of protein molar mass based on gel electrophoresis experiments was performed.

3. Results and Discussion

Effect of Triton® X-114 concentration.

The Figure 1 shows the partition coefficients (K_p) as a function of Triton® X-114 concentration (2-15 % w/v). Variations in the K_p values (0.21 ± 0.02 and 0.57 ± 0.05 for 2 and 15 % w/v Triton® X-114, respectively) were observed when the Triton® X-114 concentration increased by a factor of ca. seven. Considering the difficulty to manipulate a 15 % (w/v) Triton® X-114 solution (the higher K_p value observed) due to its high viscosity, and due to the K_p value remained almost constant between 5 and 10 % (w/v), a 5 % (w/v) Triton® X-114 concentration was then used as the optimal condition.

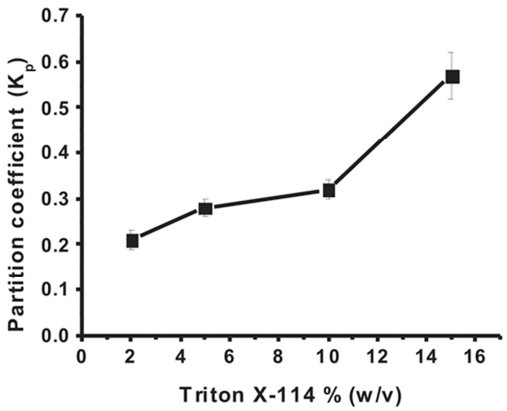


Figure 1. Partition coefficient as a function of Triton® X-114 concentrations using 2 mL of sample volume, 10 % (w/v) NaCl, 20 min of contact time and pH 7.2.

Proportion of volume between sample and surfactant solution.

According to the Figure 1, the partition coefficient increased when surfactant concentrations were raised. However, as commented before, the highly viscous surfactant solution is difficult to handle, while 5 % (w/v) Triton® X-114 was possibly not enough to extract higher quantities of protein contained in a certain volume of the human plasma. In the Figure 1, for example, the results were obtained using 2 mL of sample volume. Thus, the total protein mass was changed, ranging sample volumes from 50 to 2000 μL, keeping the surfactant concentration constant. These experiments allowed determining the proportion between sample volume and solution volume of the surfactant. The proportions 1:160, 1:80, 1:40, 1:16, 1:8 and 1:4 related to sample volume: surfactant solution were evaluated. By decreasing this proportion, a decrease in the partition coefficient was observed from 0.62 ± 0.02 to 0.30 ± 0.01 when 1:80 and 1:4 were respectively used (results not shown). Protein aggregation at higher concentrations can be happen, decreasing the migration of proteins to the surfactant-rich phase. Then, 1:80 was fixed as sample volume:surfactant solution proportion.

Effect of NaCl concentration.

The effect of electrolytes on phase separation of nonionic surfactants is well known. The electrolyte salts out the polyoxyethylated surfactants by dehydration [35,36], decreasing the cloud point temperature. In this way, NaCl concentrations were evaluated from 6 to 12 % (w/v) range and two phases were visually formed at room temperature (25 °C).

The Figure 2 shows the partition coefficient (K_p) as a function of NaCl concentration. It was observed an increase in K_p as NaCl concentration was raised, a constant behavior between 10-12 % (w/v). This behavior can be explained due to the volumes of surfactant-rich phases obtained after the two phase separations. When the NaCl concentration was increased, smaller volumes of surfactant-rich phase (2.5 mL at 6 % and 1.2 mL at 12 % w/v) were observed (results not shown). For small surfactant-rich phase volumes, it was supposed that the proteins are distributed into a small volume,

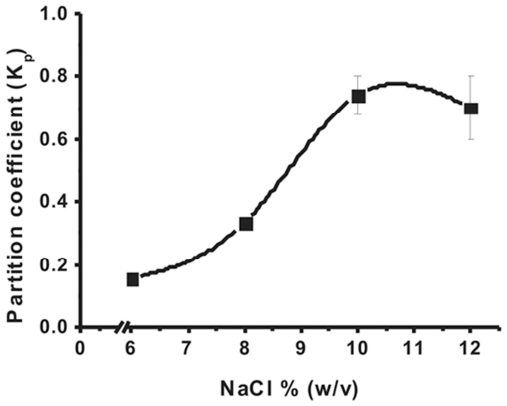


Figure 2. Partition coefficient as a function of NaCl concentration using 5 % (w/v) Triton® X-114, 100 µL of sample volume, 20 min of contact time and pH 7.2.

the preconcentration of the proteins being naturally obtained, and the partition coefficient value increased. In this context, because of a constant behavior of K_p above 10 % (w/v) NaCl, this concentration was used for subsequent separations.

Effect of contact time and pH.

After mixing all reagents, the partition coefficient remained constant from 0 to 30 min evaluated range, demonstrating that the extraction equilibrium was quickly attained. This results is according to the literature, where only slight variations in the alcohol dehydrogenase activity were found when increasing the contact time from 1 to 11 min [37], and from 10 to 50 min for albumin extraction [20]. Thus, a shorter contact time between protein and surfactant aggregates is preferable.

Few reports in the literature using the cloud point method [20,38,39] describe the influence of the pH value on protein extraction. On the other hand, most protein extraction methods are pH dependent. The pH can change the charge of chemical groups present on side chains of the amino acids, modifying the net global charge of the protein. The number of charged groups on the protein molecule surface varies by changing the pH, which implies in the modifications in protein structure [40]. This behavior also changing surfactant micelles and protein interactions and, consequently, the partition coefficient.

Saitoh and Hinze [41] reported hydrophobic membrane protein extraction over pH 5.5 to 6.5 using a zwitterionic surfactant C9-APSO4; however, hydrophilic proteins were not extracted at this pH range. In the present work, the effect of pH was then evaluated over the physiological pH range (6.4 – 7.4) with the goal in maintaining the protein structure. The partition coefficients ranged from 0.49 ± 0.08 for pH 6.4 to 0.33 ± 0.05 for pH 7.4, respectively. Because the small variation of K_p in the pH range studied, the process was carried out at pH 6.8 (K_p = 0.48 ± 0.04).

PAGE profile and ApoA-I identification.

Figure 3 shows the bands of the proteins present in the surfactant-rich and -poor phases, after the optimized extraction by cloud point (5 % w/v Triton® X-114, 100 µL of sample volume, 10 % w/v NaCl, contact time < 1 min and pH 6.8). According to this figure, surfactant-rich phase did not present protein bands at molar masses higher than 100 kDa (see lane 3), indicating a poor interaction between the surfactant micelles and these proteins. Two bands (lane 3) were observed in the surfactant-rich phase. The R2 band may be attributed to albumin (ca. 66 kDa), which presents high concentration in human plasma (51-71 % of total protein) [42], contributing to its extraction to this phase. Even considering that albumin presents remarkable hydrophilic characteristics, its presence in the surfactant-rich phase is inherent to CPE since this phase is somewhat hydrated. The R1 band (ca. 27 kDa) is attributed to the ApoA-I protein, and the densitometric analysis (N=9) reveled that this band presents ca. 174 ± 6 ng of protein. Additionally, this protein can also be observed in the surfactant-poor phase (lane 4).

Mass spectrometry analyses were performed to confirm the identity of ApoA-I present in surfactant-rich phase after separation using two-dimensional gel electrophoresis. In this case, the success on protein characterization depends on the adopted separation strategy. In this way, the use of a multidimensional separation system avoids the overlap of different protein bands in the same region of the gel and allows adequate correlation with proteins contained in protein databanks.

Two-dimensional electrophoretic profile obtained for the proteins contained in surfactant rich-phase is shown in Figure 4. Analysis of spot A allowed identification of three peptides, described in Table 1, which had their primary sequence

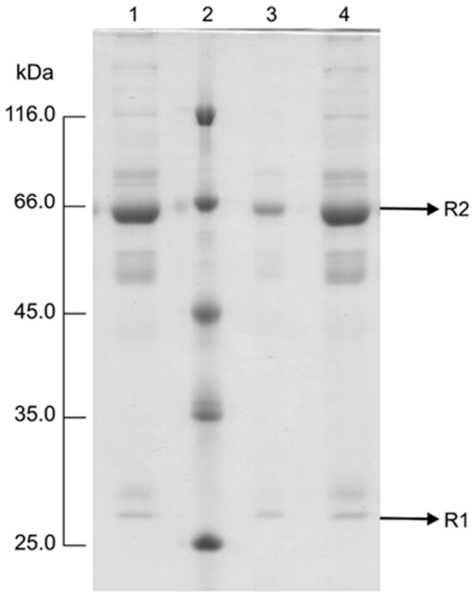


Figure 3. SDS-PAGE of proteins fractions obtained with the optimized conditions using the cloud point extraction. Lane 1 = plasma matrix without phase separation; Lane 2 = standard molar mass markers; Lane 3 = surfactant-rich phase (5.2 µg); Lane 4 = surfactant-poor phase (10.5 µg). Albumin and ApolipoproteinA-I are represented by R1 and R2, respectively.

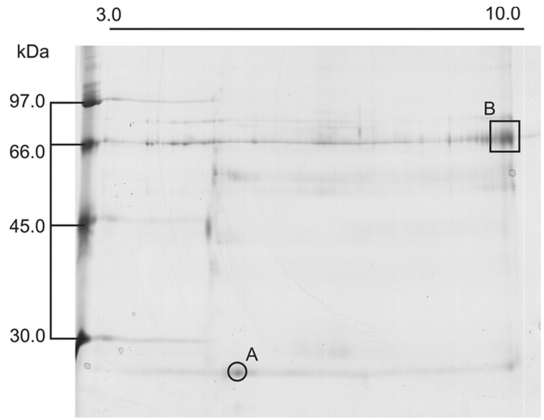


Figure 4. Separation of the proteins present in the surfactant-rich phase by two-dimensional gel electrophoresis. Spot in region A indicates ApoA-I protein, while region B represents albumin.

revealed by tandem mass spectrometry, through spectra shown in Figure 5. In fact, those three identified peptides correlate with ApoA-I protein (access code CAA00975), correspond to a coverage of 12% of the primary structure of the protein (with a MASCOT score of 97). The theoretical mass (28061 Da) obtained from MSDB databank is in accordance to the experimental value obtained after 2D-PAGE separation (ca. 26650 Da), indicating that the variation between theoretical and experimental masses is lower than 5%.

4. Concluding remarks

An alternative method for ApoA-I extraction from human plasma based on a cloud point strategy was developed. Biocompatible conditions were used for ApoA-I separation, such as room temperature and pH 6.8, with 5 % (w/v) Triton® X-114 and 10 % (w/v) NaCl. At optimized conditions, the partition coefficient (ca. 0.5), calculated from the total protein concentration, indicates a relatively low affinity of the human plasma proteins and surfactant aggregates. However, the electrophoretic separation by 2D-PAGE related to those proteins in the surfactant-rich phase presented a clear gel, identifying the presence of the target protein (ApoA-I), which was successfully extracted to this phase. The band of ApoA-I was also observed in the surfactant-poor phase through SDS-PAGE analysis. The small amount of the surfactant present in the poor phase can explain this behavior.

The similarity of this procedure with that one previously published [14], demonstrates its robustness, what should be, in our opinion, the target of any analytical proteomic approach.

Table 1. ApoA-I peptides identified by tandem mass spectrometry.

Experimental m/z value	Experimental molecular mass (Da)	Theoretical molecular mass (Da)	Peptide
1012.5754	1011.5681	1011.5713	K.AKPALEDLR.Q
1226.5383	1225.5310	1225.5364	-.DEPPQSPWDR.V
1301.6343	1300.6270	1300.6411	R.THLAPYSDELR.Q

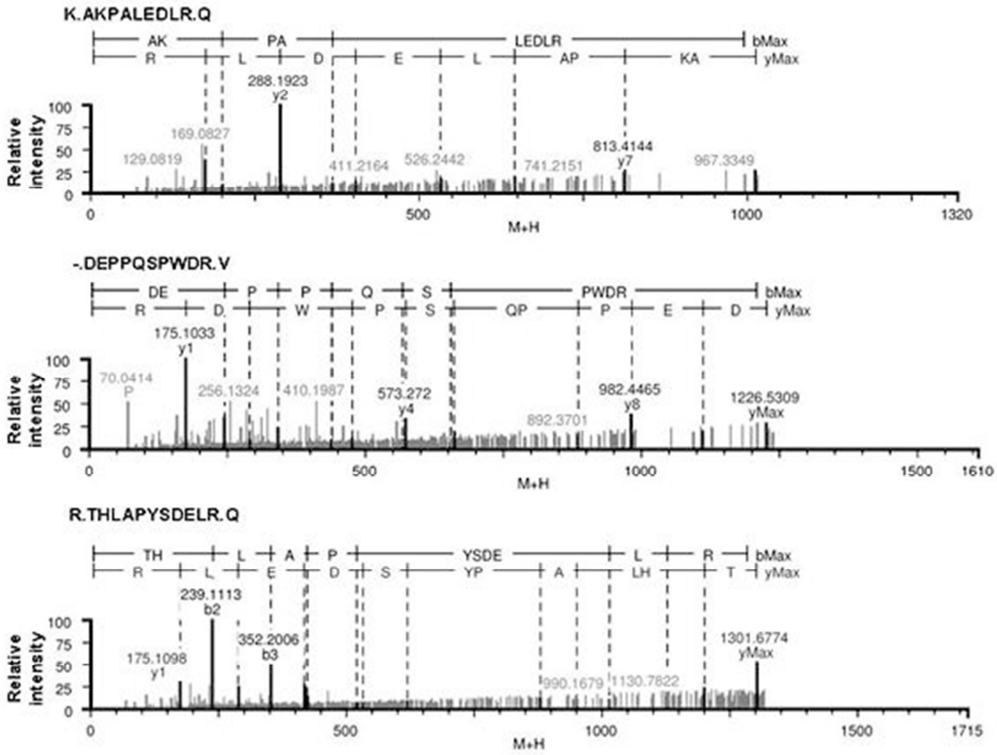


Figure 5. Fragment ion spectra and fragmentation profile obtained for identified peptides shown in Table 1.

Finally, the procedure can easily be scaled-up, employing a simple, fast (ca. 15 min) and low cost method (US$ < 0.50 per sample) for ApoA-I separation. Additionally, it can be considered a less aggressive process compared to traditional methodologies due to the use of nonionic surfactants only and simple electrolytes.

Acknowledgements

The authors thank the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP, São Paulo, Brazil), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brasília, Brazil), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brasília, Brazil) and the Financiadora de Estudos e Projetos (FINEP, Brasília, Brazil) for financial support and fellowships. The authors also thank to the Brazilian Synchrotron Light Laboratory – Associação Brasileira de Tecnologia de Luz Síncrotron (Proteome Network of the São Paulo State – FAPESP process number 2004/14846-0 / and FINEP process number 01.07.0290.00) for use of the mass spectrometry laboratory facilities and Carol H. Collins for language assistance.

References

1. C.B. Ojeda, F.S. Rojas, Anal. Bioanal. Chem. 394 (2009) 759-782.
2. Z.L. Wang, Appl. Microbiol. Biotechnol. 75 (2007) 1-10.
3. H.F. Cheng, D.A. Sabatini, (2007) Sep. Sci. Technol., 42 (3) 453-475.
4. R. Carabias-Martínez, E. Rodrígues-Gonzalo, B. Moreno-Cordero, J.L. Pérez-Pavón, C. García-Pinto, E.F. Laespada, J. Chromatogr. A, 902 (2000): 251-265.
5. L.E.G. Vara, B.L. Alfaro, Anal. Biochem. 387 (2009) 280-286.
6. C.O. Rangel-Yagui, H. Lam, D.T. Kamei, D.I.C. Wang, A. Pessoa Jr., D. Blankschtein, Biotechnol. Bioeng. 82 (2003) 445-456.
7. C.S. Magalhães, J.S. Garcia, A.S. Lopes, E.C. Figueiredo, M.A.Z. Arruda, in: M.A.Z. Arruda (Ed.), Trends in sample preparation, Nova Science Publishers, New York, USA, 2007, pp. 245-288.
8. P.S. Monteiro, J.S.R. Coimbra, L.A. Minim, J.A. de Oliveira, L.H.M. da Silva, J. Chromatogr., B 867 (2008) 189-193.
9. L.M. Coelho, M.A. Bezerra, M.A.Z. Arruda, R.E. Bruns, S.L.C. Ferreira, Sep. Sci. Technol. 43 (2008) 815-827.
10. W.L. Hinze, E. Pramauro, Crit. Rev. Anal. Chem. 24 (1993) 133-177.
11. Y.J. Nikas, C.L. Liu, T. Srivastava, N.L. Abbott, D. Blankschtein, Macromoleculles 25 (1992) 4797-4806.
12. K. Selber, F. Tjerneld, A. Collén, T. Hyytiä, T. Nakari-Setälä, M. Bailey, R. Fergerström, J. Kan, J. Van Der Lann, M. Penttilä, M.R. Kula, Process Biochem. 39 (2004) 889-896.
13. H. Lam, M. Kavoosi, C.A. Haynes, D.I.C. Wang, D. Blankschtein, Biotechnol. Bioeng. 89 (2005) 381-392.
14. A.S.Lopes, J.S. Garcia, R.R. Catharino, L.S. Santos, M.N. Eberlin, M.A.Z. Arruda, Anal. Chim. Acta 590 (2007) 166-172.
15. M.D. Rukhadze, S.K. Tsagareli, N.S. Sidamonidze, V.R. Meyer Anal. Biochem. 287 (2000) 279-283.
16. S. Saito, In: M.J. Schick (Ed.) Nonionic surfactant: physical chemistry, Marcel Dekker, New York, USA, 1987, pp. 881-926.
17. C. Bordier, J. Biol. Chem. 256 (1981) 1604-1607.
18. S.R. Sirimanne, D.G. Patterson Jr., L. Ma, J.B. Justice Jr., J. Chromatogr., B 716 (1998) 129-137.
19. B. Yao, L. Yang, Sep. Sci. Technol. 43 (2008) 1564-1580.
20. T. Minuth, H. Gieren, U. Pape, H.C. Raths, J. Thömmes, M.R. Kula Biotechnol. Bioeng. 55 (1997) 339-347.
21. M.B. Linder, M. Qiao, F. Laumen, K. Selber, T. Hyytiä, T. Nakari-Setälä, M.E. Penttilä, Biochemistry 43 (2004) 11873-11882.
22. D. Roosmalen, M.P.J. Dohmen-Speelmans, C.H.J.T. Dietz, L.J.P. Van Den Broeke, L.A.M. Van Der Wielen, J.T.F. Keurentjes, Food Bioprod. Process. 84 (2006) 51-58.
23. J.S. Becker, O.R.T. Thomas, M. Franzreb, Sep. Purif. Technol. 65 (2009) 46-53.
24. P.S. Monteiro, J.S.R. Coimbra, L.A. Minim, J.A. Oliveira, L.H.M. Silva J. Chromatogr. B 867 (2008) 189-193.
25. M.A.O. da Silva, M.A.Z. Arruda Talanta 77 (2009) 985-990.
26. K. Björhall, T. Miliotis, P. Davidsson, Proteomics, 5 (2005) 307-317.
27. B.A. Chromy, A.D. Gonzales, J. Perkins, M.W. Choi, M.H. Corzett, B.C. Chang, C.H. Corzett, S.L. McCutchen-Maloney, J. Proteom. Res. 3 (2004) 1120-1127.
28. U. Beisiegel, Eur. Herat J. 19 (1998) A20-A23.
29. J. Persson, L. Nyström, H. Ageland, F. Tjerneld, J. Chromatogr. B 711 (1998) 97-109.
30. M.M. Bradford, Anal. Biochem. 72 (1976) 248-254.
31. J.S. Garcia, G.H.M.F. Souza, M.N. Eberlin, M.A.Z. Arruda, Metallomics 1 (2009) 107-113.
32. G. Candiano, M. Bruschi, L. Musante, L. Santucci, G.M. Ghiggeri, B. Carnemolla, P. Orecchia, L. Zardi, P.G. Righetti, Electrophoresis 25 (2004) 1327-1333.
33. P. Onnerfjord, S. Ekstrom, J. Bergquist, J. Nilsson, T. Laurell, G. Marko-Varga, Rapid Commun Mass Spectrom. 13 (1999) 315-322.
34. D.N. Perkins, D.J.C. Pappin, D.M. Creasy, J.S. Cottrell, Electrophoresis 20 (1999) 3551-3567.
35. T. Iwanaga, M. Suzuki, H. Kunieda, Langmuir 14 (1998) 5775-5781.
36. H. Schott J. Colloid Interface Sci. 173 (1995) 265-277.
37. T.X. Zhang, H.Z. Liu, J.Y. Chen J. Chem. Technol. Biotechnol. 75 (2000) 798-802.
38. Sánchez-Ferrer, M. Pérez-Gilabert, E. Núñez, R. Bru, F. García-Carmona, J. Chromatogr. A 668 (1994) 75-83.
39. K. Selber, A. Collén, T. Hyytiä, M. Penttilä, F. Tjerneld, M.R. Kula, Bioseparation 10 (2001) 229-236.
40. J. Gutenwik, B. Nilsson, A. Axelsson, Biochem. Eng. J. 19 (2004) 1-7.
41. T. Saito, W.L. Hinze, Anal. Chem. 63 (1991) 2520-2525.
42. R.W. Putnam, The plasma proteins: structure, function, and genetic control, second ed., Academic Press, New York, USA, 1975.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.26

Analysis of Histone Exchange during Chromatin Purification

Stephanie Byrum[1], Samuel G. Mackintosh[1], Ricky D. Edmondson[1], Wang L. Cheung[1], Sean D. Taverna[2], and Alan J. Tackett*[1].

[1]University of Arkansas for Medical Sciences, 4301 West Markham Street, Little Rock, Arkansas 72205, USA; [2]Johns Hopkins School of Medicine, 725 North Wolfe Street, Baltimore, Maryland 21287, USA.

Received: 12 July 2010 **Accepted:** 28 August 2010 **Available Online**: 10 September 2010

ABSTRACT

Central to the study of chromosome biology are techniques that permit the purification of small chromatin sections for analysis of associated DNA and proteins, including histones. Chromatin purification protocols vary greatly in the extent of chemical cross-linking used to prevent protein dissociation/re-association during isolation. Particularly for genome-wide analyses, chromatin purification requires a balanced level of fixation that captures native protein-protein and protein/DNA interactions, yet leaving chromatin sections soluble and accessible to affinity reagents. We have applied a relative quantification methodology called I-DIRT (isotopic differentiation of interactions as random or targeted) for optimizing levels of chemical cross-linking for affinity purification of cognate chromatin sections. We show that fine-tuning of chemical cross-linking is necessary for isolation of chromatin sections when minimal histone/protein exchange is required.

Keywords: Cross-linking; Histone; Chromatin; Affinity Purification.

Abbreviations

I-DIRT, isotopic differentiation of interactions as random or targeted; **MALDI,** matrix-assisted laser desorption ionization; **FA**, formaldehyde; **ChIP,** chromatin immunoprecipitation; **ChIP-chip**, chromatin immunoprecipitation with DNA chip readout; **ChIP-seq**,chromatin immunoprecipitation with DNA sequencing readout.

1. Introduction

The eukaryotic genome is highly organized into transcriptionally active or repressive chromatin compartments, which consist of repeating octamers of histones called nucleosomes. Access to these regions may be epigenetically regulated in part by covalent post-translational modifications (PTMs) of histone proteins [1]. Histone PTMs are proposed to act as chemical flags that functionally partition chromatin through direct binding/targeting of protein complexes with distinct properties [2]. The field of chromatin biology employs technologies like ChIP (chromatin immunoprecipitation), affinity purification of protein/histone complexes for proteomic analysis, and more recent technology that allows for the purification of chromosome sections for proteomic analysis [3-5]. Central to each of these techniques is the purification of chromatin sections with cognate histones. To overcome the inherent exchange of histones and other proteins during the isolation of chromatin sections, investigators utilize *in vivo* chemical cross-linking with agents such as formaldehyde. However, a quantitative analysis of the level of protein exchange has not been reported. Additionally, the purification of a chromatin bound protein complex can be challenging as too much cross-linking renders the complex insoluble, while too little cross-linking does not trap less stable protein interactions [6].

Here we utilize an isotopic labeling approach with affinity purification to readily gauge levels of histone exchange in purified chromatin samples. The approach described is an application of our previously reported I-DIRT (isotopic differentiation of interactions as random or targeted) technology (Fig. 1) [7]. The fundamental basis of I-DIRT is the mixing

***Corresponding author:** Alan J. Tackett. University of Arkansas for Medical Sciences, 4301 West Markham Street, Little Rock, Arkansas 72205, USA. Tel. 501-686-8152; fax 501-686-8169; Email Address: ajtackett@uams.edu.

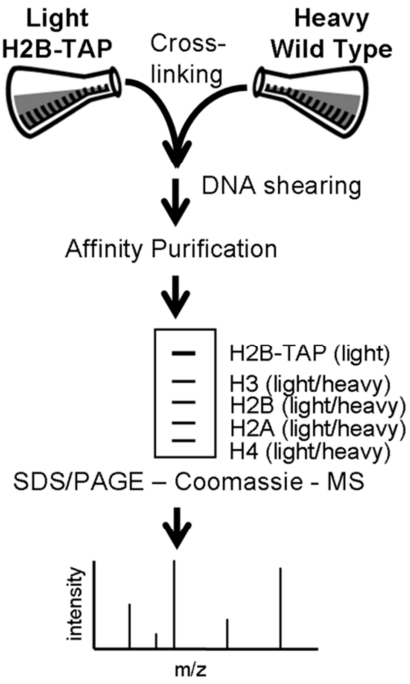


Figure 1. I-DIRT analysis of histone exchange during chromatin purification. H2B-TAP cells were grown isotopically light (12C6-Arg), while non-tagged cells were grown isotopically heavy (13C6-Arg). Cultures were treated with various levels of formaldehyde. Cells were harvested independently and mixed 1:1 for co-cryogenic lysis. Chromatin was sheared and then affinity purified on IgG coated Dynabeads. Co-purifying histones were resolved by SDS-PAGE and the ratios (isotopically light to heavy arginine containing histone peptides) were measured with high resolution mass spectrometry.

of an isotopically light, affinity tagged cell lysate with an isotopically heavy, non-tagged cell lysate – such that proteins purifying with the tagged, isotopically light protein are exclusively isotopically light, while those purifying non-specifically are a 1:1 mix of light and heavy proteins. The 1:1 mix observed for non-specifically associating proteins can be correlated to proteins that readily exchange during the time course of the affinity purification. Other approaches similar to I-DIRT have also been applied to study specific protein interactions in the presence of cross-linking [8-10]. One example of these approaches is the quantitative analysis of tandem affinity-purified *in vivo* cross-linked protein complexes (QTAX) strategy that utilizes extensive chemical cross-linking and stringent immunopurification [8]. I-DIRT and other strategies have been used to analyze functional protein complexes, but not specifically to analyze structures like chromatin. In the work reported here, we chose to use our I-DIRT strategy to follow the exchange of histones during the purification of small chromatin sections. We show that *in vivo* chemical cross-linking is necessary to prevent histone exchange during chromatin purification, and the approach presented provides the methodology to study histone exchange dynamics for techniques requiring the purification of cognate chromatin sections.

2. Material and methods

Saccharomyces cerevisiae HTB1::TAP-HIS3 BY4741 (Open Biosystems) cells were grown in isotopically light synthetic media, while an arginine auxotrophic strain (*arg4::KAN BY4741*, Open Biosystems) was grown in isotopically heavy synthetic media ($^{13}C_6$ arginine, 80 mg/L, Cambridge Isotope Laboratories, CLM-2265). Synthetic media consisted of 6.7 g/L yeast nitrogen base without amino acids (Sigma), 2 g/L synthetic drop-out media minus lysine (US Biological), 80 mg/L lysine (Fisher) and 20% (w/v) glucose (Fisher). Both strains were grown to ~3 x 10^7 cells/mL at 30°C, cross-linked for 5 minutes with formaldehyde (0, 0.05, 0.25 or 1.25% formaldehyde (Sigma)), and quenched for 5 minutes with 125 mM glycine. Cells were harvested, frozen as pellets in liquid nitrogen, mixed 1:1 (isotopically light cells: heavy cells) by cell weight, and co-cryogenically lysed with a Retsch MM301 mixer mill. One gram of each lysate (equivalent to ~1.5 x 10^{10} cells) was re-suspended in 5 mL of affinity purification buffer (20 mM HEPES pH 7.4, 300 mM NaCl, 0.1% tween-20, 2 mM $MgCl_2$, and 1% Sigma fungal protease inhibitors). Chromosomal DNA was sheared to ~800nt sections with a Bioruptor (Diagenode). The Bioruptor was set to 12 cycles of 30 seconds with sonication followed by 30 seconds without sonication, set to the "high" sonication option, and maintained at 4°C with a circulating water bath. The resulting lysates were clarified by centrifugation (2,500 x g) for 10 min. H2B-TAP was collected from the supernatants with 4 mg of IgG-coated Dynabeads (Invitrogen) for 4 hours at 4°C [6]. Beads were washed 5-times with affinity purification buffer and treated with 0.5 N ammonium hydroxide / 0.5 mM EDTA to elute proteins. Eluted proteins were lyophilized, re-suspended in a reducing SDS-PAGE loading buffer and heated at 90°C for 20 min (which provided for reversal of formaldehyde cross-links).

Proteins were resolved on 4-20% Novex Tris-Glycine gels (Invitrogen), visualized by colloidal Coomassie staining and the region of the gel containing core histones was excised as 2-mm bands for protein identification. Proteins were digested in-gel with 100 ng trypsin (Roche) and peptides were subjected to tandem mass spectrometric analysis with a coupled Eksigent NanoLC-2D and Thermo LTQ-Orbitrap mass spectrometer [6]. Briefly, peptides were eluted from a New Objective IntegraFrit column (10 cm, 50 μm ID) packed with 4 μm Phenomenex Jupiter Proteo resin over 50 minutes with a gradient of 0.1% acetonitrile / 0.1% formic acid to 75% acetonitrile / 0.1% formic acid (0.5 μL/min flow rate) and fragmented at 35% collision energy. Precursor ions were measured in the Orbitrap mass analyzer, while peptide fragmentation and fragment ion detection occurred in the linear ion trap. A Mascot (version 2.2.03) database search identified isotopically light and heavy arginine containing histone peptides. Database parameters were as follows: precursor ion tolerance 10 ppm, fragment ion tolerance 0.6 Da, fixed modification of carbamidomethyl on cysteine, variable modification of oxidation on methionine, and 2 missed cleavages possible with trypsin (see Supplemental Table 1 for a list of all

proteins identified). Monoisotopic peak areas were extracted for each of the arginine containing histone peptides from the raw data files using Qual Browser version 2.0 (Thermo) and the percent isotopically light peptide was calculated (Supplemental Table 2). For each histone reported, multiple arginine-containing peptides were identified and the average percent isotopically light is reported.

3. Results and Discussion

We investigated the utility of *in vivo* chemical cross-linking with formaldehyde for preventing histone exchange during chromatin purification. Chemical cross-linking has been utilized to trap *in vivo* protein-protein and/or protein-DNA interactions for mass spectrometric analysis; however, extensive cross-linking can render the chromatin insoluble [6]. To identify the amount of chemical cross-linking required to prevent histone exchange during chromatin purification, the relative quantitative affinity purification strategy I-DIRT (isotopic differentiation of interactions as random or targeted) was utilized (Figure 1) [7]. *Saccharomyces cerevisiae* cells containing a TAP-tagged histone H2B were grown in isotopically light synthetic media, while an arginine auxotrophic strain (*arg4::KAN*) was grown in isotopically heavy synthetic media ($^{13}C_6$ arginine). Both cultures were independently cross-linked with formaldehyde; and harvested cells were mixed 1:1 (isotopically light cells: heavy cells) for co-cryogenic lysis. Chromosomal DNA was sheared to ~800nt sections (Figure 2A). As the percentage of cross-linking increased the ability to shear genomic DNA decreased as shown in Figure 2A. Low levels of sheared DNA were detectable up to 1.25% formaldehyde cross-linking. The purification of these sheared chromosomal sections via H2B-TAP was followed by western blotting (Figure 2B). As observed for DNA shearing in Figure 2A, increasing amounts of formaldehyde cross-linking showed a decrease in the ability to purify sheared chromatin. A cross-linking level of 1.25% formaldehyde was found to be near the upper limit of chemical cross-linking that still provided for shearing and enriching chromatin sections.

Chromatin sections were next purified from ~1.5 x 1010 cells (1:1 mixture of light H2B-TAP and heavy non-tagged cells) and proteins co-purifying with H2B-TAP were resolved by SDS-PAGE (Figure 2C). H2B-TAP, H2B, H2A, H3 and H4 were visible by Coomassie staining and detected by mass spectrometry at each level of formaldehyde cross-linking tested. Proteins were digested in-gel with trypsin and peptides were subjected to tandem mass spectrometric analysis with a coupled Eksigent NanoLC-2D and Thermo LTQ-Orbitrap mass spectrometer [6]. A Mascot database search identified arginine containing histone peptides. Monoisotopic peak areas were extracted for each of the arginine containing histone peptides and the percent isotopically light peptide was calculated. The following numbers of arginine-containing histone peptides were used for the percent isotopically light calculation: histone H2B-TAP (100 peptides), H3 (169 peptides), H2A (45 peptides), and H4 (95 peptides). If

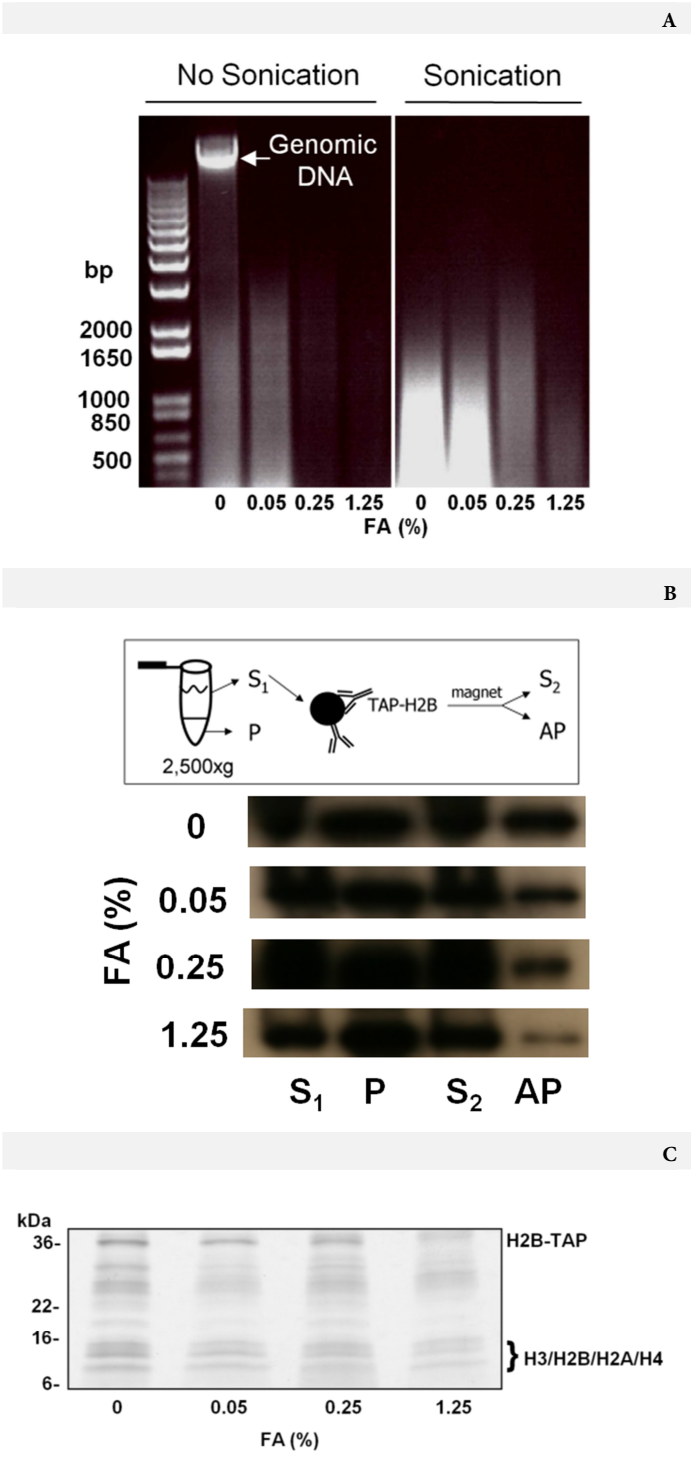


Figure 2. Purification of formaldehyde cross-linked chromatin. (A) DNA shearing as a function of formaldehyde (FA) cross-linking. DNA was isolated from formaldehyde treated H2B-TAP cells, resolved by electrophoresis and visualized by ethidium bromide staining. (B) Affinity purification of chromatin was monitored by western blotting for H2B-TAP. P, lysate pellet; S1, pre-purification supernatant; S2, post-purification supernatant; AP, affinity purified. (C) Sheared chromatin was affinity purified then histones were resolved by SDS-PAGE, visualized by Coomassie staining and excised for mass spectrometric analysis/identification.

an arginine-containing histone peptide was identified more than once in the analysis, then each spectrum was used in the calculation of percent isotopically light peptides. Representative mass spectra for arginine containing peptides from H2B-TAP and H4 are shown in Figure 3A & B. Reported and averaged together in Figure 3C are both unmodified and post-translational modification containing histone peptides from triplicate experiments. Levels of light peptides near 100% reflect minimal exchange of a protein during chromatin purification, while levels near 50% reflect rapid exchange. In Figure 3C, we show the average of ten non-specifically associating proteins, which are often observed as contaminants in affinity purifications (Rp59b, L39B, Ssb1, Rp52, Ssc1, Act1, Eno1, Hsc82, Ssa2, Fba1). The percent isotopically light for these non-specific proteins was approximately 50-60% - reflecting rapid exchange of these non-specifically associating proteins with the purified chromatin. Arginine containing peptides from H2B-TAP were exclusively light because the tagged protein was only expressed in the isotopically light culture (Figure 3A & C). Histone H2A showed minimal exchange with H2B-TAP, which indicates that the H2B/H2A interaction is stable with or without cross-linking. However, histone H3 and H4 showed a similar level of exchange (~10-20%) from 0.05% to 0.25% formaldehyde cross-linking (Figure 3C). This level of H3/H4 exchange would significantly alter the purity of chromatin sections isolated for experimental analyses such as chromatin immuno-precipitation. The similar level of exchange of H3 and H4 likely reflects the preferential affinity of these histones, which form the tetrameric core of the nucleosome. At 1.25% cross-linking, the histone H3 and H4 showed no exchange. These results indicate that 1.25% formaldehyde cross-linking is necessary to eliminate any histone exchange during chromatin purification. It is noted that the 1.25% formaldehyde is specific for yeast synthetic media, as other medias will require different levels of cross-linking in accordance to their amine or cross-linking moiety content. Thus, at 1.25% formaldehyde cross-linking in yeast synthetic media, soluble chromatin sections can be generated by sonication and isolated by affinity purification without histone exchange.

We present the application of I-DIRT technology for determining the level of histone dissociation/re-association during purification of chromatin. The technique will be broadly applicable for demonstrating purification of cognate chromatin sections in different cellular growth medias and with various *in vivo* chemical cross-linking reagents. Determining the optimal level of *in vivo* chemical cross-linking is needed to find the balance between solublizing the chromatin and preventing histone exchange. This greater emphasizes the need to optimize the level of *in vivo* cross-linking as one desires to prevent exchange while maximizing purification.

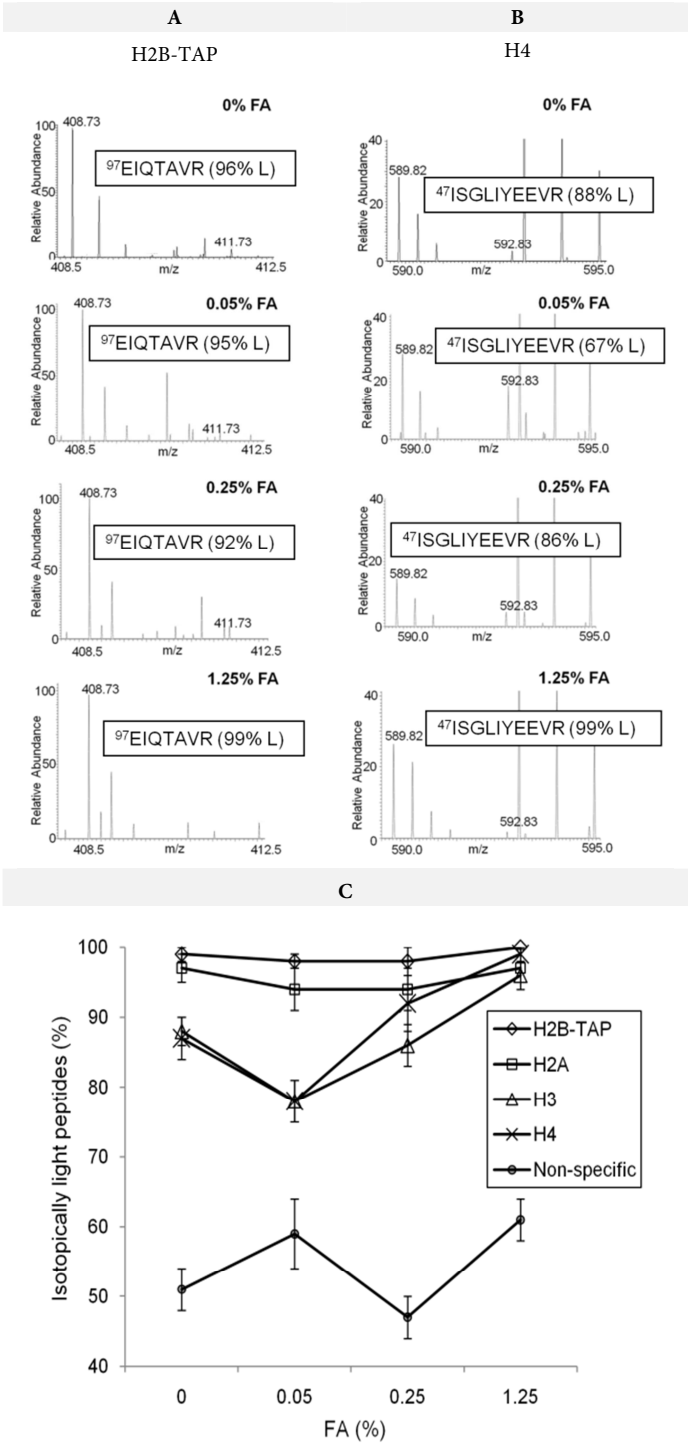


Figure 3. In vivo chemical cross-linking prevents histone exchange of cross-linked chromatin. Shown are representative mass spectra collected with an Orbitrap mass analyzer for doubly charged peptides from histone H2B-TAP (A) and histone H4 (B). The percent isotopically light peptide (%L) is indicated. (C) Percentage isotopically light arginine containing histone peptides are reported as a function of formaldehyde (FA) cross-linking. The standard error from triplicate experiments is shown. Levels approaching ~100% indicate minimal exchange, while those at ~50% reflect rapid exchange. The average of ten rapidly exchanging / non-specific proteins is shown (non-specific).

4. Future perspectives

Currently, the localization of proteins and histone PTMs on chromosomes is largely studied with techniques like ChIP-chip and more recently ChIP-seq. These technologies

are quite powerful for the high resolution localization of a given protein or modified histone; however, they lack the ability to simultaneously identify all proteins bound and the combinatorial nature of the modified histones at a given chromosomal region. As the field matures, technology development will move toward the isolation of sections of chromatin for mass spectrometric analysis of cognate histones and bound proteins. Recent examples of this are the isolation of telomeric and origin of replication chromatin for mass spectrometric analyses [5,11]. The technique we report here will play a significant role in these types of studies as one would need to ensure that the purified chromatin is representative of the *in vivo* setting.

5. Supplementary material

Supplementary material regarding this manuscript is online available in the web page of JIOMICS.

http://www.jiomics.com/index.php/jio/rt/suppFiles/26/39

Acknowledgements

This work was funded by NIH R01DA025755 (AJT & SDT) and KL2RR029883 (WLC).

References

1. C.D. Allis, T. Jenuwein, D. Reinberg, M.L. Caparros, Epigenetics, Cold Spring Harbor Laboratory Press (2006)
2. S.D. Taverna, H. Li, A.J. Ruthenburg, C.D. Allis, D.J. Patel, Nat Struct Mol Biol 14 (2007) 1025-40.
3. A.J. Tackett, D.J. Dilworth, M.J. Davey, M. O'Donnell, J.D. Aitchison, M.P.Rout, B.T.Chait, J Cell Biol 169 (2005) 35-47.
4. D.K. Pokholok, C.T. Harbison, S. Levine, M. Cole, N.M. Hannett, T.I. Lee, G.W. Bell, K. Walker, P.A. Rolfe, E. Herbolsheimer, J. Zeitlinger, F. Lewitter, D.K. Gifford, R.A. Young, Cell 122 (2005) 517-27.
5. J. Dejardin, R.E. Kingston, Cell 136 (2009) 175-86.
6. S.K. Smart, S.G. Mackintosh, R.D. Edmondson, S.D. Taverna, A.J. Tackett, Protein Sci 18 (2009) 1987-97.
7. A.J. Tackett, J.A. DeGrasse, M.D. Sekedat, M. Oeffinger, M.P. Rout, B.T. Chait, J Proteome Res 4 (2005) 1752-6.
8. C. Guerrero, C. Tagwerker, P. Kaiser, L. Huang, Mol Cell Proteomics 5 (2006) 366-78.
9. D.F. Tardiff, K.C. Abruzzi, M. Rosbash, Proc Natl Acad Sci U S A 104 (2007) 19948-53.
10. C. Tagwerker, K. Flick, M. Cui, C. Guerrero, Y. Dou, B. Auer, P. Baldi, L. Huang, P. Kaiser, Mol Cell Proteomics 5 (2006) 737-48.
11. A. Unnikrishnan, P.R. Gafken, T. Tsukiyama, Nat Struct Mol Biol 17 (2010) 430-437.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.28

A simulated annealing-based algorithm for iterative class discovery using fuzzy logic for informative gene selection

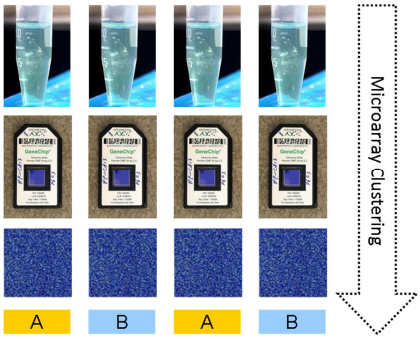
Daniel Glez-Peña[1], Miguel Reboiro-Jato[1], Florentino Fdez-Riverola*[1], Fernando Díaz[2].

[1]Dept. Informática, University of Vigo, Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense, Spain. Email Address: dgpena@uvigo.es, mrjato@sing.ei.uvigo.es, riverola@uvigo.es; [2]Dept. Informática, University of Valladolid, Escuela Universitaria de Informática, Plaza Santa Eulalia, 9-11, 40005, Segovia, Spain. Email Address: fdiaz@infor.uva.es

Received: 21 July 2010 **Accepted:** 28 August 2010 **Available Online**: 9 September 2010

ABSTRACT

Within a gene expression matrix, there are usually several particular macroscopic phenotypes of samples related to some diseases or drug effects, such as diseased samples, normal samples or drug treated samples. The goal of sample-based clustering is to find the phenotype structures or sub-structure of these samples. We present a novel method for automatically discovering clusters of samples which are coherent from a genetic point of view. Each possible cluster is characterized by a fuzzy pattern which maintains a fuzzy discretization of relevant gene expression values. Possible clusters are randomly constructed and iteratively refined by following a probabilistic search and an optimization schema. Evaluation of the proposed algorithm on publicly available microarray datasets shows high accuracy in spite of noise and the presence of other clusters. The results obtained support the appropriateness of using fuzzy logic to represent and filter gene expression values following an iterative approach. The proposed method complements our previous GENECBR system and both are freely available under GNU General Public License from http://www.genecbr.org/fpclustering.htm and http://www.genecbr.org/, respectively.



Keywords: Microarray data; Fuzzy discretization; Gene selection; Simulated annealing.

1. Introduction

Following the advent of high-throughput microarray technology it is now possible to simultaneously monitor the expression levels of thousands of genes during important biological processes and across collections of related samples. In this context, sample-based clustering is one of the most common methods for discovering disease subtypes as well as unknown taxonomies. By revealing hidden structures in microarray data, cluster analysis can potentially lead to more tailored therapies for patients as well as better diagnostic procedures.

From a practical point of view, existing sample-based clustering methods can be (*i*) directly applied to cluster samples using all the genes as features (i.e., classical techniques such as K-means, SOM, HC, etc.) or (*ii*) executed after a set of informative genes are identified. The problem with the first approach is the signal-to-noise ratio (smaller than 1:10), which is known to seriously reduce the accuracy of clustering results due to the existence of noise and outliers of the samples [1]. To overcome such difficulties, particular methods can be applied to identify informative genes and reduce gene dimensionality prior to clustering samples in order to detect their phenotypes. In this context, both supervised and unsupervised informative gene selection techniques have been developed.

While supervised informative gene selection techniques often yield high clustering accuracy rates, unsupervised

***Corresponding author:** Florentino Fdez-Riverola. Dept. Informática, University of Vigo, Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense, Spain. Email Address: riverola@uvigo.es.

informative gene selection methods are more complex because they assume no a priori phenotype information being assigned to any sample [2]. In such a situation, two general strategies have been adopted to address the lack of prior knowledge: (*i*) unsupervised gene selection, this aims to reduce the number of genes before clustering samples by using appropriate statistical models [3-5] and (*ii*) interrelated clustering, that takes advantage of utilizing the relationship between the genes and samples to perform gene selection and sample clustering simultaneously in an iterative paradigm.

Following the second strategy for unsupervised informative gene selection (interrelated clustering), Ben-Dor *et al.* [6] present an approach based on statistically scoring candidate partitions according to the overabundance of genes that separate the different classes. Xing and Karp [1] use a feature filtering procedure for ranking features according to their intrinsic discriminability and irredundancy to other relevant features. Their clustering algorithm is based on the concept of a normalized cut for grouping samples in new reference partition. von Heydebreck *et al.* [7] and Tang *et al.* [8] propose algorithms for selecting sample partitions and corresponding gene sets by defining an indicator of partition quality and a search procedure to maximize this parameter. Varma and Simon [9] describe an algorithm for automatically detecting clusters of samples that are discernable only in a subset of genes. They use iteration between Minimal Spanning Tree based clustering and feature selection to remove noise genes in a step-wise manner while simultaneously sharpening the clustering.

In this article we improve a previous initial work [10] by providing *i*) a complete mathematical formulation of the proposed method, *ii*) an evaluation of our method using two real datasets, herein referred as HC-Salamanca dataset [11] and Armstrong dataset [12] (see Sections 3.1 and 3.2 for a detailed description of these datasets), and *iii*) a comparison of the results obtained by the proposed method against the ones obtained by the standard hierarchical clustering algorithm for the same datasets. As introduced in [10], our clustering technique is based on the notion of *genetic coherence* of the each cluster, and this "coherence" is computed by taking into consideration the genes which share the same expression value through all the samples belonging to the cluster (which we term a *fuzzy pattern* or FP in short), but discarding those genes present due to pure chance (herein referred to *noisy genes* of a fuzzy pattern). The proposed clustering technique combines both (*i*) the simplicity and good performance of a heuristic search method able to find good partitions in the space of all possible partitions of the set of samples with (*ii*) the robustness of fuzzy logic, able to cope with several levels of uncertainty and imprecision by using partial truth values.

2. Material and methods

2.1 Overview of the proposed method.

As mentioned above we propose a simulated annealing-based algorithm for iterative class discovery. It uses a novel fuzzy logic method for informative gene selection. The interrelated clustering process carried out is based on an iterative approach in which possible clusters are randomly constructed and evaluated by following a probabilistic search and an optimization schema. Our clustering technique is not based on the distance between the microarrays belonging to each given cluster, but on the notion of genetic coherence of the own clusters. The genetic coherence of a given partition is calculated by taking into consideration the genes which share the same expression value through all the samples belonging to the cluster (we term this a fuzzy pattern), but discarding those genes present purely by chance (or noisy genes of a fuzzy pattern). The global view of the proposed method is sketched in Figure 2 and following sections give details about the mathematical background and proposed algorithm.

2.2 Fuzzy discretization and fuzzy pattern construction.

A fuzzy pattern is based on the fuzzy discretization given by three membership functions which are associated with each probe set in the microarray. Basically, for each probe set we consider three linguistic labels (Low, Medium and High levels of gene expression), each one associated with a polynomic function. Given a fixed value for the θ parameter (θ defines the threshold for the membership function from which a linguistic label is activated), the different labels are only activated in specific intervals within the whole range of

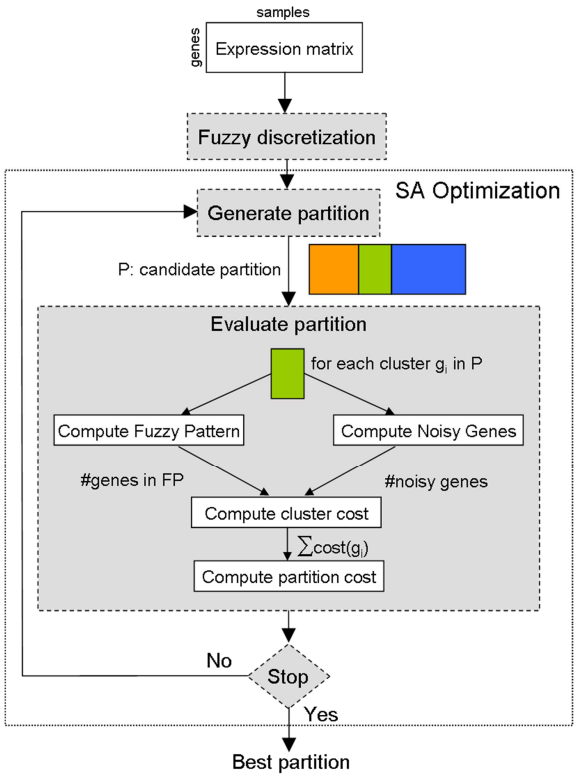


Figure 2. Overview of the iterative class discovery method. This figure shows how, from the fuzzy discretization of the microarrays from raw dataset, the method performs a stochastic search, looking for a "good partition" of microarrays in order to maximize the genetic coherence of each one cluster within the tentative partition. The simulated annealing technique is used to implement this search.

variation of a gene's expression level as shown in Figure 3.

For a specific probe set and a given value of the θ parameter, the label assigned to the gene's expression level in a probe will be one of the following alternatives:

- one of the three basic labels (LOW, MEDIUM and HIGH), if the numeric value is in only one of the associated intervals,
- one of the two combined labels (LOW-MEDIUM or MEDIUM-HIGH), if the numeric value is at the intersection of two intersections,
- the empty label, if the numeric value is in an interval where none of the labels are activated for the selected θ parameter.

Therefore, for each gene probe set we are considering a universe of six possible symbols: 'Low' (L), 'Low-Medium' (L-M), 'Medium' (M), 'Medium-High' (M-H), 'High' (H) and '*' (empty), herein denoted by $\boldsymbol{S} = \{s1, s2, s3, s4, s5, s6\}$ where sj maps the j-th symbol in the list given above.

Once the discretization of the whole microarray dataset $\boldsymbol{D}$ has been completed, given a subset of m microarrays, $D_m = \{x1, x2, \ldots, x_m\} \subseteq \mathbf{D}$, which represents any target concept (a class within a classification or a group in a clustering), its associated fuzzy pattern is constructed by selecting those linguistic labels which are different to the empty label and have a relative appearance frequency in set $\mathbf{D}_m$ equal to or greater than the predefined ratio given by the π parameter (with $0 < \pi \leq 1$). Formally, for a specific gene g_j ($1 \leq j \leq N$, where N is the number of probes in the microarray), the appearance frequency of any symbol $s \in \boldsymbol{S}$ in the set $\boldsymbol{D}_m$, $freq_j(s)$, can be computed according to the following expression:

$$freq_j(s) = \frac{\sum_{1 \leq i \leq m} \delta_j(\boldsymbol{x}_i, s)}{m}, \text{ where } \delta_j(\boldsymbol{x}_i, s) = \begin{cases} 1 & \text{if DiscretizedLabel}(\boldsymbol{x}_{ij}) = s \\ 0 & \text{otherwise} \end{cases}$$

Then, the gene g_j (with the label associated with the most frequent symbol sm_f) is selected as a candidate for the fuzzy pattern only if the most frequent symbol is different from the empty label and its appearance frequency is at least equal to the π parameter, namely, $freq_j(sm_f) \geq \pi$. Therefore, the parameter π controls the degree of exigency for selecting a gene in the fuzzy pattern, since the higher the value of the π parameter the fewer the number of genes which form the fuzzy pattern associated to the target concept $\boldsymbol{D}_m$.

Basically, our assertion is that a fuzzy pattern is able to capture the meaningful genes of any group of microarrays which are coherent from a genetic point of view. The underlying hypothesis is that any subtype of a given disease must necessarily have an internal genetic coherence, namely, those microarrays belonging to patients which suffer from one specific subtype, should share a large number of genes (i.e., present a similar expression level in a large number of genes, at least, more than if the microarrays come from patients with different subtypes of the same or a different disease). This fact has been empirically observed in previous experiments carried out with our GENECBR platform, a translational tool for multiple-microarray analysis and integrative information retrieval for aiding diagnosis in cancer research, when we studied the differences between the number of genes in the FP belonging to well and not well-defined pathologies [13-15]. This circumstance supports the development of new approaches able to take advantage of this situation. The underlying idea is that if this behaviour is observed in well-known and well-characterized classes of a disease, it must also be true in unknown groups representing new subtypes of the disease. Therefore, these newly discovered classes must be characterized by having a large number of genes in their associated fuzzy patterns. As a consequence, this situation can be used to consider the problem of clustering microarrays in terms of maximizing the number of genes in the fuzzy pattern associated with each cluster.

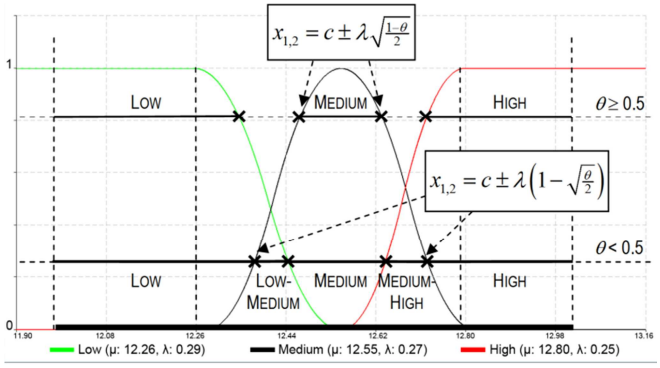


Figure 3. Membership functions and activation regions. This figure illustrates the three membership functions for a specific probe and shows the cut points which determine the length of the segment of each fuzzy expression level within the probe domain.

A key issue is the setting-up process of the parameters θ and π, since the computation of fuzzy patterns is high sensitive to these values. Although for different learning tasks (prediction or supervised classification), in previous works [13-14], a cross-valitation strategy was used to set up these parameters for the same two datasets tested in this work and then, those values has been also used in the experimental section of this work.

2.3 Noisy genes identification.

Now, working with a set of m microarrays, we are interested in the estimation of the probability that a specific gene, g_i, of the n available in a microarray, may appear in a fuzzy pattern merely by chance. First of all, we need an estimation of the occurrence probability of each symbol in the fuzzy discretization of numeric data, namely, $p(\boldsymbol{S}) = (p(s_1), p(s_2), p(s_3), p(s_4), p(s_5), p(s_6))$. Given a fixed θ, these probabilities can be estimated by the ratio of the length of each interval (associated to the labels) and the length of the whole variation range, $\Delta = c_H - c_L + \lambda_H + \lambda_L$ (see Figure 3).

Since the membership functions are polynomic, the length of each interval can be computed in a closed form. For example, the cut points of the membership function for the 'Medium' label with the line $\mu_M(x) = \theta$, is given by:

$$x_{1,2} = c_M \pm \delta_M(\theta)$$

Where:

$$\delta_M(\theta) = \begin{cases} \lambda_M \sqrt{\frac{1-\theta}{2}} & \text{if } 0.5 \le \theta \le 1 \\ \lambda_M \left(1 - \sqrt{\frac{\theta}{2}}\right) & \text{if } 0 \le \theta < 0.5 \end{cases}$$

In this way, the probabilities of each symbol in $\boldsymbol{S}$ can be computed, and obviously, their sum is equal to 1.

In a second step, having an estimate of the probability of each possible symbol, we need to assess the probability that the gene g_i appears in the fuzzy pattern associated with a sample of m microarrays with a minimum frequency ratio equal to π. Under these conditions, in order to include the gene in the fuzzy pattern it is necessary that in the set of m microarrays, it must have at least $\lceil \pi \cdot m \rceil$ repetitions of the same symbol in their associated fuzzy representations. The empty label (with symbol '*') must be excluded since it reflects that none of the labels are activated and therefore, it never can be part of a fuzzy pattern. By $p(k)$ we denote the probability that a valid label (all except the '*' symbol), appears exactly in k discretized values (of the m available). It can be shown that this probability is given by:

$$\begin{aligned} p(k) &= \sum_{j=1}^{5} p(k \mid s_j) p(s_j) \\ &= \sum_{j=1}^{5} p(s_j) \cdot C_m^k \cdot \left(p(s_j)^k (1 - p(s_j))^{m-k} \right) \end{aligned}$$

Where $C_m{}^k$ is the number of ways of picking k unordered outcomes from m possibilities, and $p(s_j)^k(1 - p(s_j))^{m-k}$ stands for the probability that symbol s_j appears k times in a sequence of m symbols.

Therefore, the probability that a gene g_i appears in the fuzzy pattern which is associated with a sample of m microarrays, is given by the sum of individual probabilities that any symbol appears, at least, $\lceil \pi \cdot m \rceil$ times. This probability can be calculated by means of the following expression:

$$\begin{aligned} p(g_i) &= \sum_{k=\lceil \pi \cdot m \rceil}^{m} p(k) \\ &= \sum_{k=\lceil \pi \cdot m \rceil}^{m} \sum_{j=1}^{5} p(s_j) \cdot C_m^k \cdot \left(p(s_j)^k (1 - p(s_j))^{m-k} \right) \end{aligned}$$

Where $p(g_i)$ is upper bound by 1, being closer to 1 depending on the distribution probability of each gene and without taking account of the degree of exigency imposed by π. For example, assuming that the probability of the empty label is null, the probability is closer to one as there is a predominant symbol (a symbol with occurrence probability close to 1), whereas the worst case is represented by the situation where all the valid symbols have the same probability of occurrence.

Finally, assuming that, in the random case, the selection of a gene g_i is independent of the selection of another one g_j ($j \neq i$), the number of noisy genes due to the chance for a group of m microarrays with N probes at the levels imposed by θ and π parameters, is given by:

$$\begin{aligned} ng(m, \theta, \pi) &= \sum_{i=1}^{N} p(g_i) = \sum_{i=1}^{N} \sum_{k=\lceil \pi \cdot m \rceil}^{m} p(k) \\ &= \sum_{i=1}^{N} \sum_{k=\lceil \pi \cdot m \rceil}^{m} \sum_{j=1}^{5} p(s_j) \cdot C_m^k \cdot \left(p(s_j)^k (1 - p(s_j))^{m-k} \right) \end{aligned}$$

Where $ng(m, \theta, \pi)$ is upper bound by N, although this value is only reachable in the ideal case where only one symbol has probability 1 for all of the genes.

As the uncertainty decreases (there is a predominance of one expression level over the other ones for all the genes in the available set of microarrays) the number of $ng(m, \theta, \pi)$ decreases (the amount of information encoded by the data also decreases and then, there are more irrelevant genes). When uncertainty increases, the amount of information also grows and more genes are necessary to distinguish samples in absence of other information. Figure 4 illustrates the variation of noisy genes depending on the θ, and π parameters.

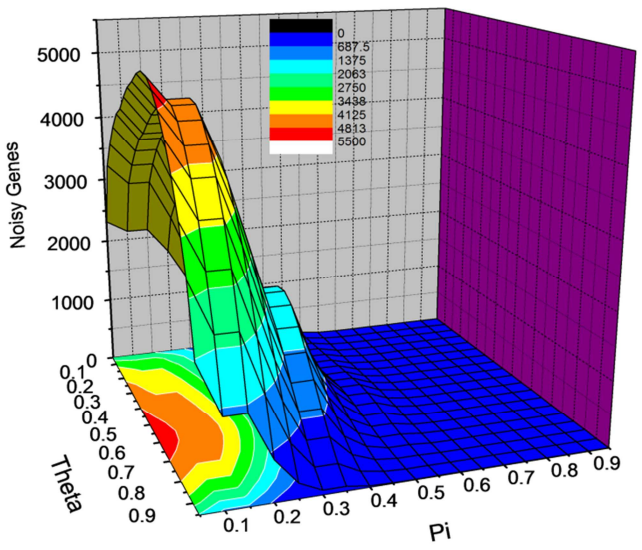


Figure 4. Noisy genes vs. Theta (θ) and Pi (π) parameters (m = 43 samples, n = 22,283 genes). This figure shows the variation in the number of noisy genes depending on the θ and π parameters of a fuzzy pattern for the HC-Salamanca dataset. For a fixed value of parameter θ, the number of noisy genes decreases exponentially when the π parameter grows (since the prior probability that a gene appears in the fuzzy pattern also decreases). For a fixed value of parameter π, the maximum number of noisy genes is symmetrically distributed around θ = 0.5 (since the prior probability of each linguistic label in a fuzzy pattern is nearly equiprobable).

2.4 The cost function of a cluster

The cost function must combine two factors: (*i*) the number of genes in the fuzzy pattern associated with each cluster of the partition and (*ii*) the size of each such cluster. The first factor in the cost function models the genetic coherence of a cluster. Accepting this hypothesis, it is expected that for clusters with equal sizes, the number of genes in a fuzzy pattern will be greater if the genetic coherence of the cluster is higher. The second factor is relevant since it has been experimentally observed that meaningful genes in large clusters (after noisy genes have been filtered out) are several orders of magnitude smaller than meaningful genes computed in small clusters. This fact is reasonable because it will be more probable when the number of possibilities is also further reduced. Therefore, the size factor in the cost function is needed for examining comparable clusters of different size.

Under these assumptions, given the available set of microarrays denoted by $\boldsymbol{X}$, and a partition $\boldsymbol{P} = \{g_1, g_2, \ldots, g_k\}$ of $\boldsymbol{X}$ in k clusters, that is to say, $g_i \subseteq \boldsymbol{X}$ with $1 \le i \le k$, $g_i \cap g_j = \varnothing$ (if $i \ne j$) and $\cup g_i = \boldsymbol{X}$, the cost of a cluster $g_i \in \boldsymbol{P}$ is given by:

$$\cos t(g_i) = \frac{N}{2^{|g_i|}\left(\left|fp(g_i, P, \theta, \pi)\right| - ng(|g_i|, \theta, \pi)\right)}$$

Where $|g_i|$ is the number of microarrays in cluster g_i, $|fp(g_i, \boldsymbol{P}, \theta, \pi)|$ is the size (in genes) of the fuzzy pattern associated with the group g_i with regard to the classification given by $\boldsymbol{P}$ and for specific values of the θ and π parameters, and $ng(|g_i|, \theta, \pi)$ is the number of noisy genes of a group with $|g_i|$ microarrays. Finally, N is the fixed number of probes in a microarray. Therefore, if one tries to minimize the defined cost of a group, $cost(g_i)$, it involves trying to maximize both (*i*) the size of the cluster and (*ii*) its genetic coherence (measured by the number of genes belonging to its fuzzy pattern).

Finally, the cost of a given partition $\boldsymbol{P} = \{g_1, g_2, \ldots, g_k\}$ of $\boldsymbol{X}$ is defined by the sum of the individual costs of each group:

$$\cos t(P) = \sum_{i=1}^{k} \cos t(g_i)$$

Once the cost of a tentative partition of microarrays has been established, one needs to define an algorithmic strategy in order to automatically build random partitions from the set of available microarrays.

2.5 Algorithm

There are certain optimization problems that become unmanageable using combinatorial methods as the number of objects becomes large. The simulated annealing technique which can be regarded as a variant of a local search was first introduced by Metropolis *et al.* [16], and then used in optimization problems by Kirkpatrick *et al.* [17] and Černy [18]. For these problems, the simulated annealing method represents a very effective practical algorithm. Although this technique is unlikely to find the optimum solution, it can often discover a very good one even in the presence of noisy data.

Simulated annealing improves its behaviour through the introduction of two tricks. The first one is the so-called *Metropolis algorithm* [16], in which some poor solutions are accepted when they serve to allow the solver to explore more of the possible solution space. Such bad solutions are tolerated using the criterion that:

$$e^{-\Delta D/T} > U(0, 1)$$

Where ΔD is the variation of the cost function for the current solution and the best one, T stands for a *synthetic temperature* and $U(0, 1)$ is a random number in the interval [0, 1]. The cost function D corresponds to the free energy in the case of annealing a metal (in which case the temperature parameter would actually be the kT, where k is Boltzmann's Constant and T is the physical temperature in the Kelvin absolute temperature scale). If T is large, many bad solutions are accepted and a considerable part of the solution space is accessed. The next solutions to explore are randomly constructed, though more sophisticated techniques can be used.

The second trick is, again by analogy with the annealing of a metal, to lower the *temperature*. After making many changes in the current solution and observing that the cost function declines only slowly, one lowers the temperature limiting while the size of allowed bad solutions. After lowering the temperature several times to reach a smaller value, one may then "quench" the process by accepting only good solutions in order to find the local minimum of the cost function. There are various annealing schedules for lowering the temperature, but the results are generally not very sensitive to the details. These general ideas are the basis of simulated annealing but a comprehensive introduction to the subject can be found in [19].

The application of our simulated annealing approach to cluster microarrays is sketched in Figure 5. First of all, we consider a pool which contains the set of m microarrays that must be clusterized into k different and unknown groups. In the final solution, some microarrays can stay in the pool without being associated with any cluster. Initially, a first solution to the problem (a partition of microarrays) is randomly constructed. All the microarrays of the pool are randomly distributed among k classes, where k is the desired number of clusters in the partition (the whole set of m microarrays are spread proportionally among the k clusters and the pool).

At every step, a neighbour solution is determined by choosing one from the following alternatives (see Figure 6 for the details):

- Moving a randomly chosen microarray from the pool to a cluster (perhaps empty).

- Moving a randomly chosen microarray from a cluster to the pool.
- Exchanging randomly chosen microarrays among clusters.
- Exchanging randomly chosen microarrays among a cluster and the pool.
- Moving a randomly chosen microarray from one cluster to another cluster.

The neighbour solutions of lower cost obtained in this way are always accepted, whereas those solutions with a higher cost are accepted with the following probability:

$$Pr = T_i / (T_i + \delta)$$

Where δ is the cost difference among the new solution and the old solution, and T_i (i = 0, 1, …) represents the temperature of annealing which drops from a value T_0 (the cost of the initial solution) according to the formula $T_{i+1} = T_i \cdot \alpha$, where $\alpha < 1$. Pr implies that large increases in the solution cost (uphill moves) are more likely to be accepted when T_i is high. As T_i approaches zero most uphill moves are rejected.

The general algorithm stops if equilibrium is encountered. We define that equilibrium is reached if, after 50 stages of temperature reduction, the best achieved solution can not be improved. In contrast to the classical approach in which a solution to the problem is taken as the last solution obtained in the annealing process, we memorize the best solution found during the whole annealing process (Cf. lines 13-15 in Figure 5). Moreover, at the beginning of each temperature epoch, the search is restarted from the best solution reached at the moment (Cf. line 6 in Figure 5).

Summing up, the proposed annealing algorithm performs the local search by sampling the neighbourhood randomly. It attempts to avoid becoming prematurely trapped in a local optimum by sometimes accepting low-grade solutions. The acceptance level depends on the magnitude of the increment of the solution cost and on the spent search time. By this reason, and specially, at initial stages, when the temperature

```
   Input:
   → microarray dataset (pool of m microarrays) to be grouped in an unsupervised way
   → number of clusters (k)
   Output:
   ← partition of the original dataset into k clusters
   Require:
   next_solution routine
   cost function
   Steps:
1  current_solution = initial partition {Builds randomly a partition of k clusters with the microarrays in the pool}
2  best_solution = current_solution {initialize the best partition built at the moment}
3  equilibrium_counter = 0 {initialize the counter which controls the annealing epochs without improvement of the best found solution}
4  T = cost(current_solution) {initial temperature of the annealing process}
5  repeat
6      current_solution = best_solution {The annealing epoch starts from the best partition at the moment}
7      for iteration_counter = 1 to m do {An annealing epoch is made up of m attempts, where m is the number of microarrays in the pool}
8          new_solution = next_solution(current_solution) {Builds a new partition from the current one}
9          δ = cost(new_solution) – cost (current_solution) {Computes the difference in cost of the new partition and the current partition}
10         x = u(0, 1) {Generate random x uniformly in the range (0, 1)}
11         if ( δ < 0 ) or ( x < T / (T+δ) ) then {Accept a new solution if it improves the cost or increases the cost but it has a high probability of acceptance, the term T /(T+δ) which depends on the current temperature and the difference of the costs}
12             current_solution = new_solution {update the current solution}
13             if ( cost(new_solution) < cost (best_solution) ) then {if new partition improves best partition at the moment }
14                 best_solution = new_solution {update best partition}
15                 equilibrium_counter = 0 {reinitialize the equilibrium counter}
16     T = T · α {Decrease current temperature multiplying by a constant rate α = 0.95}
17     equilibrium_counter = equilibrium_counter + 1 {increment the number of epochs without improvement}
18 until equilibrium_counter > 50 {stop the annealing process when a stationary state is reached, at least 50 epochs without improvement}
```

Figure 5. General pseudo code of simulated annealing-based clustering algorithm. This algorithm explains the steps involved in partitioning a microarray dataset into *k* clusters by grouping microarrays which maximize its genetic coherence (assessed in terms of the number of genes in their associated fuzzy pattern), using a simulated annealing search algorithm.

Input:
→ current partition of the original microarray dataset (current_solution)
Output:
← neighbour partition which is built from the input current_solution by randomly choosing one of five possible movements of microarrays between clusters
Require:
-

Steps:

```
1   new_solution = current_solution {The new partition is built from the current partition}
2   Choose randomly two different clusters of new_solution, c_i and c_j
3   Select randomly three microarrays: m_i, m_j and m_k, belonging to c_i, c_j and the pool, respectively
4   u01 = u(0, 1) {Generate random u01 uniformly in the range (0, 1)}
5   if ( u01 < 0.2 ) then
6           move microarray m_i (from cluster c_i) to the pool
7   if ( 0.2 ≤ u01 < 0.4 ) then
8           move microarray m_k (from pool) to cluster c_i
9   if ( 0.4 ≤ u01 < 0.6 ) then
10          exchange microarray m_i (from cluster c_i) with microarray m_j (from cluster c_j)
11  if ( 0.6 ≤ u01 < 0.8 ) then
12          exchange microarray m_i (from cluster c_i) with microarray m_k (from pool)
13  if ( 0.8 ≤ u01 ≤ 1 ) then
14          move microarray m_i (from cluster c_i) to cluster c_j
15  return new_solution
```

Figure 6. Pseudo code of new_solution function. This algorithm explains the steps involved in building a neighbour partition from the current partition by randomly choosing one of the five possible operations.

is high, it has no sense to set the initial solution with a "reasonable" solution (for example, computed by a simple clustering algorithm) because the algorithm perhaps will be accept other solutions quite different from the original one, since the goal of the algorithm is to perform a global exploration of the search space. Only when the search process progresses, the exploitative component of the algorithm dominates over the explorative component, performing a local search around the selected local optimum after these intial stages. Obviously, the convergence time of the the proposed algorithm is higher than other deterministic clustering algorithms, but these algorithms have no capability to escape from local optima. Moreover the computational effor of the proposed clustering technique (up to several hours of running time per execution) since the evaluation of the cost function requires the computation of a fuzzy pattern for each cluster in the current partition.

3. Resutls and Discussion

3.1 The HC-Salamanca dataset

This dataset consists of bone marrow samples from 43 adult patients with de novo diagnosed acute myeloid leukemia (AML) – 10 acute promyelocytic leukemias (APL) with t(15;17), 4 AML with inv(16), 7 monocytic leukemias and 22 nonmonocytic leukemias, according to the WHO classification. All samples contained more than 80% blast cells and they were analyzed using high-density oligonucleotide microarrays (specifically, the Affymetrix GeneChip Human Genome U133A Array) [11]. In [11], hierarchical clustering analysis segregated APL, AML with inv(16), monocytic leukemias and the remaining AML into separate groups, so we consider this partition as the reference classification for validating our proposed technique in the following experimentation. The results of the proposed algorithm with this dataset are depicted in Figure 7.

Figure 7 shows for each available microarray the percentage of the times it has been grouped together with other microarrays belonging to the reference groups (APL, AML with inversion, Monocytic and Other AML) in the ten executions of the whole algorithm.

As can be seen in Figure 7, each sample has a different percentage of membership to each one of the reference groups. From this representation it can also be seen that the APL group is the most promising cluster since the algorithm has clustered together (in an unsupervised way) the majority of samples from this class in its ten executions. This result is consistent with the fact that this pathology is the best characterized class among the AML subtypes and, therefore, there is a high probability that microarrays within this subtype are well labelled in the reference classification. The Other-AML category seems to be another class, at least different from other clusters except the AML with-inversion group. This is the uncertain subtype of AML, since it contains those samples which are not classified within other groups. In the same way that the Other-AML group (but to a lesser degree), the monocytic leukemias seem to be another possible group. Finally, the AML with inversion is the most doubtful class since samples from this group are misclassified among the Other-AML and monocytic groups. This fact can be due

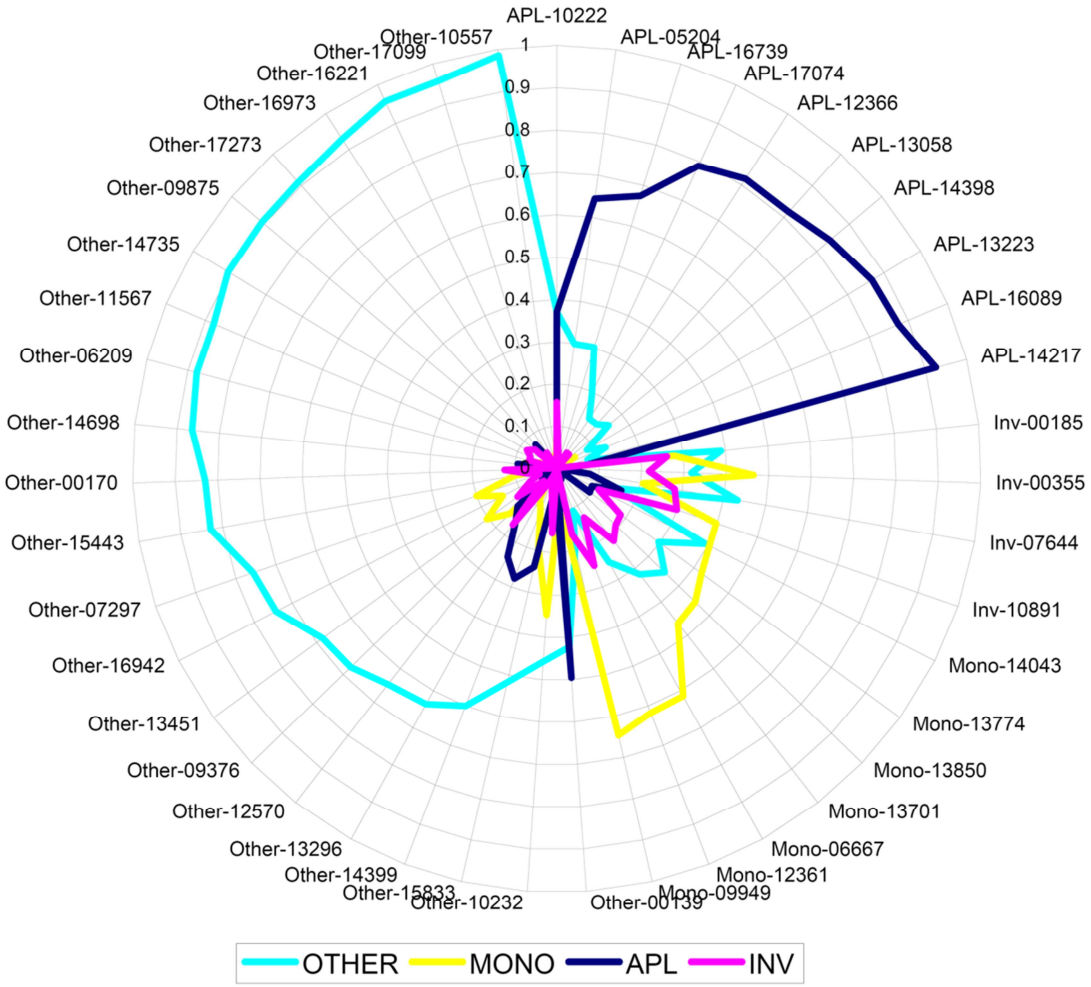


Figure 7. Degree of affinity of microarrays from the HC-Salamanca dataset with regard to the reference groups. The radial chart shows the same number of graphs as groups in the reference clustering. Each graph indicates the percentage of times in which a given microarray has been clustered with other microarrays in a reference cluster. As can be observed, the APL and Other-AML clusters are clearly differentiated, only the samples Other-10232 and APL-10222 have the highest percentage less than 50%, whereas Other-00139 is mainly grouped with samples in the APL group. The monocytic group is identified to a lesser degree but can still be differentiated. Samples from the AML with inversion group are confused with those belonging to the monocytic and Other-AML groups. Results are coherent with the hierarchical clustering reported in [11].

to the reduced number of available samples or to the lack of genetic coherence within this group, since the classification of these samples was performed by examining the karyotype by an expert.

The percentage of times (on average) in which microarrays of each reference cluster have been grouped together with microarrays belonging to different classes is shown in each row of Table 1.

Table 1 can be interpreted as a confusion matrix

Table 1. Confusion matrix for the HC-Salamanca dataset.

		Predicted class			
		APL	*Inv*	*Mono*	*Other*
True class	*APL*	76.19%	2.71%	2.18%	18.92%
	Inv	7.79%	26.49%	33.66%	32.06%
	Mono	3.11%	17.81%	51.73%	27.35%
	Other	8.62%	5.56%	8.70%	77.12%

numerically supporting the facts commented above, since the APL and Other-AML groups are the better identified pathologies (in an average percentage of 76.19% and 77.12% for all their samples and runs of the algorithm), followed by the monocytic leukemias (with an average percentage of 51.73%). As mentioned above, the AML with-inversion group is confused in a mean percentage of 33.66% and 32.06% with samples from monocytic and Other-AML groups, respectively.

If we consider that the highest percentage for each microarray determines the cluster to which it belongs, the final clustering obtained by our simulated annealing-based algorithm is shown in Table 2.

Assuming as "ground truth" the clustering given by experts, the performance of the clustering process can be tested by comparing the results given in both tables.

Some commonly used indices such as the *Rand index* and the *Jaccard coefficient* [20] have been defined to measure the

degree of similarity between two partitions. For the clustering given by our experiments, the *Rand index* was 0.90 and the *Jaccard coefficient* was 0.77.

Table 2. Final clustering for the HC-Salamanca dataset

APL	APL-05204, APL-10222, APL-12366, APL-13058, APL-13223, APL-14217, APL-14398, APL-16089, APL-16739, APL-17074, **Other-00139**
Mono	**Inv-00355, Inv-10891**, Mono-06667, Mono-09949, Mono-12361, Mono-13701, Mono-13774, Mono-13850, Mono-14043
Other	**Inv-00185, Inv-07644**, Other-00170, Other-06209, Other-07297, Other-09376, Other-09875, Other-10232, Other-10557, Other-11567, Other-12570, Other-13296, Other-13451, Other-14399, Other-14698, Other-14735, Other-15443, Other-15833, Other-16221, Other-16942, Other-16973, Other-17099, Other-17273

In order to compare with other standard methods, a new clustering has been computed using the hierarchical clustering algorithm (with $k = 8$, an average linkage strategy, and without any low-variance filter). In order to compute the *Rand index* and *Jaccard coefficient* with regard to this reference partition, a final partition has been computed assigning to each sample in a cluster the reference label which is the most frequent in the cluster. This final clustering obtained from the original results of the hierarchical clustering algorithm for the HC-Salamanca dataset is shown in Table 3.

Table 3. Final clustering computed from the results of the hierarchical clustering algorithm for the HC-Salamanca dataset.

APL	APL-05204, APL-10222, APL-12366, APL-13058, APL-13223, APL-14217, APL-14398, APL-16089, APL-16739, APL-17074, **Other-00139**
Mono	**Inv-00185, Inv-00355, Inv-07644, Inv-10891,** Mono-06667, Mono-09949, Mono-12361, Mono-13701, Mono-13774, Mono-13850, Mono-14043, **Other-10232, Other-10557, Other-13451, Other-15443, Other-15833**
Other	Other-00170, Other-06209, Other-07297, Other-09376, Other-09875, Other-11567, Other-12570, Other-13296, Other-14399, Other-14698, Other-14735, Other-16221, Other-16942, Other-16973, Other-17099, Other-17273

In this case, the computed *Rand index* is 0.79 (against 0.90 of our proposal) and the *Jaccard coefficient* is 0.53 (against 0.77). Besides the worse behavior of the hierarchical clustering algorithm from a quantitative viewpoint (derived from the values of these two indexes), it can be observed from Table 3 that hierarchical clustering algorithm and our algorithm agree that the Other-00139 sample is very similar to the APL samples, but hierarchical clustering algorithm does not distinguish clearly among samples from monocyte, with-inversion and other groups.

3.2 The Armstrong dataset

In [12] the authors proposed that lymphoblastic leukemias with MLL translocations (mixed-lineage leukemia) constitute a distinct disease, denoted as MLL, and show that the differences in gene expression are robust enough to classify leukemias correctly as MLL, acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). The public dataset of this work, herein referred to as the Armstrong dataset, has been also used to test our proposal. The complete group of samples consists of 24 patients with B-Precursor ALL (ALL), 20 patients with MLL rearranged B-precursor ALL (MLL) and 28 patients with acute myeloid leukemia (AML). All the samples were analyzed using the Affymetrix GeneChip U95a which contains 12600 known genes.

The results of the proposed clustering algorithm working with this dataset are shown in Figure 8. As in the previous examples, Figure 8 shows the percentage of times that each available microarray has been grouped together with other microarrays belonging to the reference groups (ALL, AML and MLL) in the ten executions of the algorithm.

The percentage of times (on average) in which microarrays of each reference cluster have been grouped together with microarrays of different classes is shown in Table 4. These percentages can be considered as an estimation of the overlapping area of the membership functions of any two potential groups in the sector associated to a true class.

As can be seen in Figure 8 (by analyzing the overlapping areas of membership graphs in the associated sectors to each one of the reference groups) the AML group is clearly distinguished from the ALL and the MLL groups (the confusion with regard to the ALL group is marginal being slightly larger with regard to the MLL group). The ALL group is clearly differentiated from the AML group. The main component of confusion in this group is from the MLL group (showing a clear overlap of the MLL and ALL membership graphs in the sector associated with the true ALL samples). Finally, the MLL can be distinguished to a lesser extent with respect to the other groups, the sources of confusion, in decreasing order, being the ALL and AML groups, respectively.

Therefore, the above assertions based on the interpretation of Figure 8, are numerically supported by the results shown in Table 4.

As in the HC-Salamanca dataset, if the highest percentage for each sample determines the cluster of the microarray, the final clustering obtained by our simulated annealing-based algorithm is shown in Table 5.

As in the previous experiment, assuming the clustering given by experts is the "ground truth", the performance of the clustering process can be examined by comparing the results given in both tables. In this case, the *Rand index* and the *Jaccard coefficient* for experiments carried out are 0.89 and

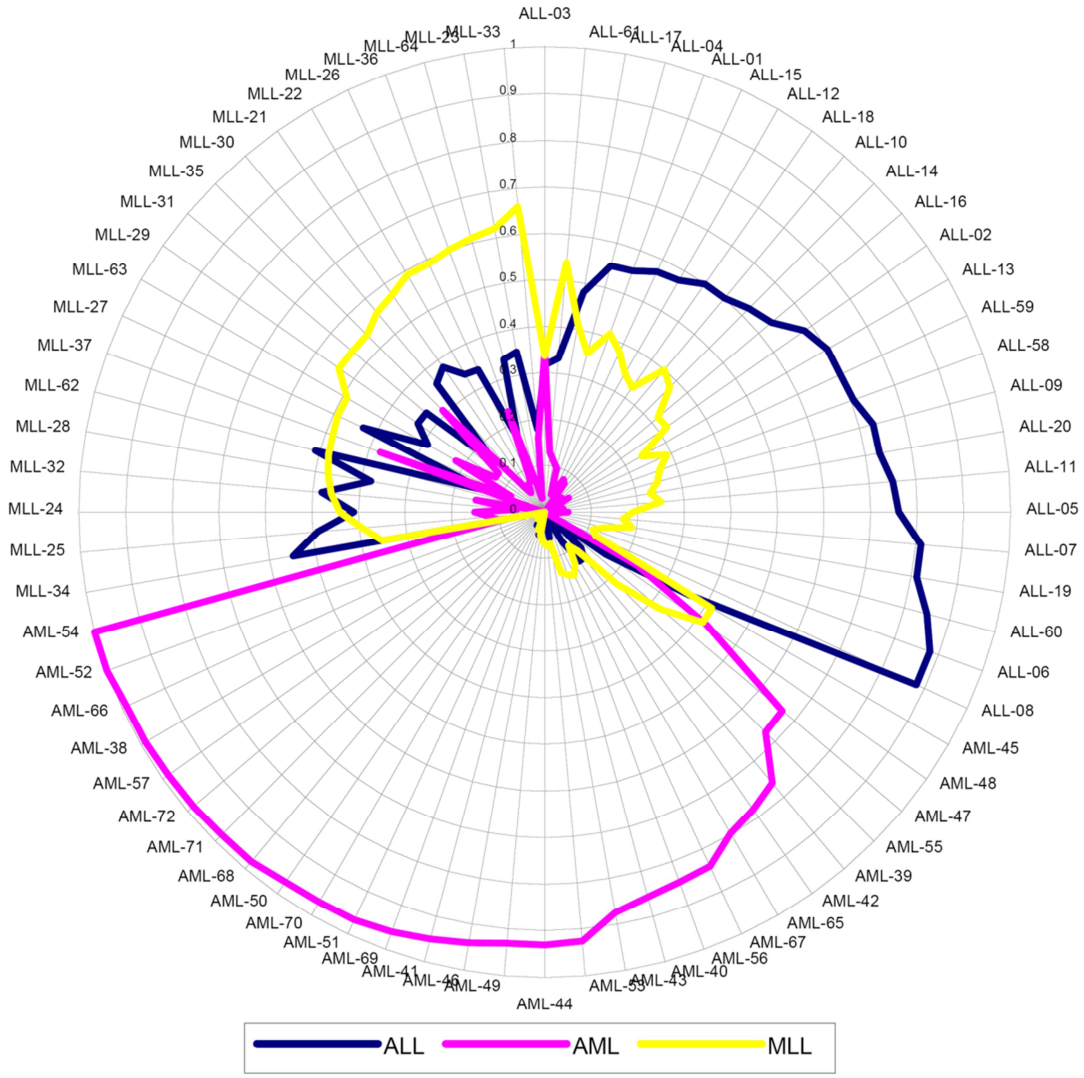


Figure 8. Degree of affinity microarrays from the Armstrong dataset with regard to the reference groups. The radial chart shows three membership graphs (one for each reference group) in the domain of available microarrays in the Armstrong dataset. From this figure the specific samples which are usually grouped with microarrays in other reference groups can be identified: the AML-45 and ALL-61 samples in the group of the MLL samples, the ALL-03 sample (the most doubtful) in the AML and MLL groups, and the MLL-25, MLL-32, MLL-34 and MLL-62 samples in the ALL group.

0.72, respectively. These indices are comparable to those obtained with the HC-Salamanca dataset. Moreover, if we assume a binary classification (considering the ALL and MLL groups as a unique partition of the lymphoblastic leukemias, the results are improved to 0.95 for the *Rand index* and 0.90 for the *Jaccard coefficient*, which are very close to the perfect match in a situation where the reference groups are also completely consistent.

Table 4. Confusion matrix for the Armstrong dataset.

		Predicted class		
		ALL	***AML***	***MLL***
True class	***ALL***	65.88%	5.16%	28.95%
	AML	4.42%	86.40%	9.18%
	MLL	34.74%	12.85%	52.41%

The hierarchical clustering algorithm (with k = 9, an average linkage strategy, and without any low-variance filter) was also executed to compute another clustering. As in previous section, from the original results of the hierarchical clustering algorithm, a final partition was built considering the most frequent reference label in a cluster to assign it to each sample in this cluster. The final clustering for the Armstrong dataset and computed from the original results of the hierarchical clustering algorithm are shown in Table 6.

In this case, the *Rand index* is 0.84 (against the 0.89 of our proposal) and the *Jaccard coefficient* is 0.64 (against 0.72), which reveal a worse behaviour of the hierarchical clustering algorithm against our clustering technique with regard to the reference partition given by the experts. From Table 6 it can be also observed that our clustering technique and the

Table 5. Final clustering for the Armstrong dataset.

ALL	ALL-01, ALL-02, ALL-04, ALL-05, ALL-06, ALL-07, ALL-08, ALL-09, ALL-10, ALL-11, ALL-12, ALL-13, ALL-14, ALL-15, ALL-16, ALL-17, ALL-18, ALL-19, ALL-20, ALL-58, ALL-59, ALL-60, **MLL-25, MLL-32, MLL-34, MLL-62**
AML	**ALL-03**, AML-38, AML-39, AML-40, AML-41, AML-42, AML-43, AML-44, AML-46, AML-47, AML-48, AML-49, AML-50, AML-51, AML-52, AML-53, AML-54, AML-55, AML-56, AML-57, AML-65, AML-66, AML-67, AML-68, AML-69, AML-70, AML-71, AML-72
MLL	**ALL-61, AML-45**, MLL-21, MLL-22, MLL-23, MLL-24, MLL-26, MLL-27, MLL-28, MLL-29, MLL-30, MLL-31, MLL-33, MLL-35, MLL-36, MLL-37, MLL-63, MLL-64

hierarchical clustering algorithm have a similar behaviour for the AML group (in the case of hierarchical algorithm is perfect) but it gets worse for the MLL group (decreasing considerably the number of samples in this cluster) and the ALL group (increasing the confusion with regard to the MLL group).

3.3 Discussion

The aim of the experiments reported in the previous section is to test the validity of the proposed clustering method. Dealing with unsupervised classification, it is very difficult to test the ability of a method to perform the clustering since there is no supervision of the process. In this sense, the classification into different groups proposed by the authors in [11-12] is assumed to be the reference partition of samples in our work. This assumption may be questionable in some cases, since the reference groups are not well established. For example, in the HC-Salamanca dataset the AML with-inversion group is established by observation of the karyotype of cancer cells, but there is no other evidence (biological, genetic) suggesting that this group corresponds to a distinct disease.

Table 6. Final clustering computed from the results of the hierarchical clustering algorithm for the Armstrong dataset.

ALL	ALL-01, ALL-02, ALL-03, ALL-04, ALL-05, ALL-06, ALL-07, ALL-08, ALL-09, ALL-10, ALL-11, ALL-12, ALL-13, ALL-14, ALL-15, ALL-16, ALL-17, ALL-18, ALL-19, ALL-20, ALL-58, ALL-59, ALL-60, ALL-61, **MLL-21, MLL-22, MLL-23, MLL-24, MLL-25, MLL-26, MLL-27, MLL-28, MLL-29, MLL-31, MLL-33, MLL-34**
AML	AML-38, AML-39, AML-40, AML-41, AML-42, AML-43, AML-44, AML-45, AML-46, AML-47, AML-49, AML-50, AML-51, AML-52, AML-53, AML-54, AML-55, AML-56, AML-57, AML-65, AML-66, AML-67, AML-68, AML-69, AML-70, AML-71, AML-72
MLL	**AML-48**, MLL-30, MLL-32, MLL-35, MLL-36, MLL-37, MLL-62, MLL-63, MLL-64

Even so, the assumption of these prior partitions as reference groups is the only way to evaluate the similarity (or dissimilarity) of the results computed by the proposed method based on existing knowledge. As it turns out, there is no perfect match among the results of our proposed method and the reference partitions, but they are compatible with the current knowledge of each dataset. For example, for the HC-Salamanca dataset the better characterized groups are the APL and Other-AML groups, the worst is the AML with inversion group, and there is some confusion of the monocytic AML with the AML with-inversion and Other-AML groups. These results are compatible with the state-of-the-art discussed in [11], where the APL group is the better characterized disease (it can be considered as a distinct class), the monocytic AML is a promising disease (in [11] the authors try to show differences in gene expression of this class with regard the others), the AML with inversion in chromosome 16 is the weaker class, and the Other-AML group acts as the dumping ground for the rest of samples which are not similar enough to the other possible classes. For the Armstrong dataset, the AML group is clearly separated from the MLL and ALL groups. It is not surprising since the myeloid leukemia (AML) and lymphoblastic leukaemias (MLL and ALL) represent distinct diseases. Some confusion is present among ALL and MLL groups, but this result is compatible with the assumption (which authors test in [12]) that the MLL group is a subtype of the ALL disease.

Moreover, the results shown in Tables 1 and 4 (by rows) are an estimation of the overlapping area between the *i*-th membership graph (associated with the *i*-th predicted group) and any *j*-th membership graph (see Figures 7 and 8 for a geometrical interpretation) taking into consideration the samples in the *i*-th true cluster. Therefore, according to the affinity graphs shown in Figures 7 and 8, these percentages can be also interpreted as a measure of the similarity/dissimilarity among predicted groups.

4. Conclusions

The simulated annealing-based algorithm presented in this work is a new algorithm for iterative class discovery that uses fuzzy logic for informative gene selection. An intrinsic advantage of the proposed method is that, assuming the percentage of times in which a given microarray has been grouped with samples of other potential classes, the degree of membership of that microarray to each potential group can be deduced. This fact allows a fuzzy clustering of the available microarrays which is more suitable for the current state-of-the-art in gene expression analysis, since it will be very unlikely to state (without uncertainty) that any available microarray only belongs to a unique potential cluster. In this case, the proposed method can help to assess the degree of affinity of each microarray with potential groups and to guide the analyst in the discovery of new diseases.

In addition, the proposed method is also an unsupervised technique for gene selection when it is used in conjunction

with the concept of discriminant fuzzy pattern (DFP) introduced in [13]. Since the selected genes depend on the resulting clustering (they are the genes in the computed DFP obtained from all groups) and the clustering is obtained by maximizing the cost function (which is based on the notion of genetic coherence and assessed by the number of genes in the fuzzy pattern of each cluster), then the selected genes jointly depend on all the genes in the microarray, and the proposed method can be also considered a multivariate method for gene selection.

Finally, the proposed technique, in conjunction with our previous developed geneCBR platform, represents a more sophisticated tool which integrates three main tasks in expression analysis: clustering, gene selection and classification. In this context, all the proposed methods are non-parametric (they do not depend on assumptions about the underlying distribution of available data), unbiased with regard to the basic computational facility used to construct them (the notion of fuzzy pattern) and with the ability to manage imprecise (and hence, uncertain) information, which is implicit in available datasets in terms of degree of membership to linguistic labels (expressions levels, potential categories, etc.).

Acknowledgements

This work has been partially funded by the Spanish Ministry of Science and Innovation, the Plan E from the Spanish Government, the European Union from the ERDF (TIN2009-14057-C03-02), Xunta de Galicia for Angeles Alvariño fellowship to D.G-P., and JCyL (Government of the autonomous community of Castile and León, ref. VA100A08).

References

1. E.P. Xing, R.M. Karp, Bioinformatics 17 (2001) S306-315.
2. [2] D. Jiang D, C. Tang , A. Zhang, IEEE T Knowl Data En 16 (2004) 1370-1386.
3. O. Alter, P.O. Brown, D. Botstein, P Natl Acad Sci USA 97 (2000) 10101-10106.
4. C.H.Q. Ding, in: Proceedings of the Sixth Annual International Conference on Computational Molecular Biology, ACM, 2002, pp. 127-136.
5. K.Y. Yeung, W.L. Ruzzo, Bioinformatics 17 (2001) 763-774.
6. A. Ben-Dor, N. Friedman, Z. Yakhini, in: Proceedings of the Fifth Annual International Conference on Computational Biology, ACM, 2001, pp. 31-38.
7. A. von Heydebreck, W. Huber, A. Poustka, M. Vingron, Bioinformatics 17 (2001) S107-114.
8. C. Tang, A. Zhang, M. Ramanathan, Bioinformatics 20 (2004) 829-838.
9. S. Varma, R. Simon, BMC Bioinformatics,5 (2004) 126.
10. D. Glez-Peña, F. Díaz, J.R. Méndez, J.M. Corchado, F. Fdez-Riverola, Lecture notes in Computer Science 5518 (2009) 972-978.
11. N.C. Gutiérrez, R. López-Pérez, J.M. Hernández, I. Isidro, B. González, M. Delgado, E. Fermiñán, J.L. García, L. Vázquez, M. González, J.F. San Miguel, Leukemia 19 (2005) 402-409.
12. S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, S.J. Korsmeyer, Nat Genet 30 (2002) 41-47.
13. F. Díaz, F. Fdez-Riverola, J.M. Corchado, Comput Intell 22 (2006) 254-268.
14. F. Fdez-Riverola, F. Díaz, M.L. Borrajo, J. C. Yáñez, J.M. Corchado, in: H. Muñoz-Avila, F. Ricci (Eds.), Proceedings of the 6th International Conference on Case-Based Reasoning, Springer, 2005, pp. 191-205.
15. D. Glez-Peña, F. Díaz, J.M. Hernández, J.M. Corchado, F. Fdez-Riverola, BMC Bioinformatics 10 (2009) 187.
16. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, J Chem Phys 21 (1953) 1087-1092.
17. S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Science 220 (1983) 671-680.
18. V. Černý V, J Optimiz Theory App 45 (1985) 41-51.
19. C.R. Reeves CR, Modern heuristic techniques for combinatorial problems, McGraw-Hill , London, 1995.
20. M. Halkidi, Y. Batistakis, M. Vazirgiannis, J Intell Inf Syst 17 (2001) 107-145.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.22

Proteomic analysis of nitrogen stress-responsive proteins in two rice cultivars differing in N utilization efficiency

Chen Song[1], Fanrong Zeng[1], Wu Feibo[1], Wujun Ma[2], Guoping Zhang*[1].

[1]Agronomy Department, Huajiachi Campus, Zhejiang University, Hangzhou 310029, China. [2]Western Australian Department of Agriculture and Food; Centre for Comparative Genomics and State Agriculture Biotechnology Centre, Murdoch University and Perth, WA 6150, Australia.

Received: 15 June 2010 **Accepted:** 9 September 2010 **Available Online**: 13 September 2010

ABSTRACT

Plant nitrogen utilization efficiency (NUE) has become critical important in modern agriculture, not only for crop growth and yield but also for reducing production cost. Moreover, one of the major negative environmental impacts of agricultural activities is associated with excessive nitrogen application. Improving NUE will ensure lower level of N fertilizer usage thus reduce environmental contamination. In order to understand the NUE mechanism of rice, the largest food crop in the world, a systematic proteomic study of investigating the nitrogen stress-responsive proteins in two rice cultivars differing in NUE is conducted. Four leaf-old seedlings were treated with normal nutrition solution and N-free solution for 12 h, 3 d and 7 d. Total proteins of leaves were extracted and separated by two-dimensional gel electrophoresis. Although more than 1000 protein spots were reproducibly detected, only a very small proportion of spots showed differential expression, including 10 and 24 up-regulated, 2 and 12 down-regulated in the two cultivars Chunyou 58 and Yongyou 6, respectively. This indicates that relatively simply biochemical pathways maybe involved with NUE thus the NUE as a trait maybe efficiently manipulated. Mass spectrometry based peptide mass fingerprinting (PMF) procedure identified 31 protein spots. Six stress-induced proteins were found, including DegP2, harpin binding proteins, Heat shock-related proteins, glutathione S-transferase GSTF14, Fibrillin-like protein and Glyceraldehyde-3-phosphate dehydrogenase. Apart from the stress related proteins, the other differential proteins identified were mainly these involved in the regulation of the main leaf biological function, photosynthesis metabolism, such as Rubisco activase, RuBisCo large subunit, etc. The study also detected two novel proteins, harpin binding protein and oryzains gamma precursor. The current study reveals new insights into N stress response and theoretical bases for improving NUE of rice crop.

Keywords: Mass spectrometry, rice (*Oryza sativa* L.), nitrogen utilization efficiency, Two-dimensional gel electrophoresis.

1. Introduction

A key element in modern agriculture is the application of nitrogen fertilizer, which has dramatically increased the crop yield [1]. In order to meet the food demand of the increased world population, application of nitrogen fertilizer in the world has been increased by 10 folds in the last half century, It was predicted that the increase trend will continue in this century, from 87 million tones in 2000 to 236 million tones in 2050 [2]. Meanwhile, most of the high yield varieties of the major crops developed in the last several decades had high nitrogen demand for the realization of yield potential [3]. On the other hand, less than half of the N fertilizers applied to the field was absorbed and utilized by crops; the majority of them was lost to the atmosphere or leached into groundwater, lakes and rivers, causing increasingly severe pollutions to the environments [4]. Therefore, development of crop cultivars with high N utilization is essential for agricultural sustainability and environmental protection.

The improvement of N fertilizer utilization could be realized by enhancing the ability of N uptake and/or increasing N utilization efficiency (NUE). For the former, a crop or a special cultivar has high ability of N uptake from the soils with low N concentration, which is referred as high uptake effi-

***Corresponding author:** Guoping Zhang. Agronomy Department, Huajiachi Campus, Zhejiang University, Hangzhou 310029, China**.** Email Address: zhanggp@zju.edu.cn

ciency. For the latter, a crop or a cultivar may make the best use of N nutrient that the plant absorbed from soil for producing biomass or harvest organs, which is evaluated by grain yield or biomass production per unit nitrogen amount[1], expressed as NUEY (N utilization efficiency of yield) and NUEB (N utilization efficiency of biomass), respectively. Although NUEY is affected by many physiological processes, including nitrogen contribution to spikelet production during early panicle formation stage, and contribution to sink size by decreasing the number of degenerated spikelets and increasing hull size during the late panicle formation stage [5], the fact that higher rice yield is achieved mainly due to greater biomass production [6] provided the possibility to explore the relationship between NUEP and NUEY.

As an essential plant macronutrient, nitrogen is required for a variety of physiological processes. It comprises 1.5–2% of plant dry matter and approximately 16% of total plant protein [4]. For rice, the leaf N, about 75% of total plant N, is associated with chloroplasts, which are physiologically important in dry matter production through photosynthesis [7]. It is also an important constituent of many important compounds, including amino acids, proteins (enzymes), nucleic acids, chlorophyll and several plant hormones.

NUE is considered as the function of N in carbohydrate production, which is closely related to the C/N balance. For plants, N and C metabolism is tightly linked in most biochemical pathways, which involve in carbon fixation, nitrogen transfer and utilization etc. Although roots play a dominant role of nitrogen uptake, leaf is the major organ for carbon and nitrogen metabolism. N drives plant dry matter production through the control of both leaf area index (LAI) and leaf photosynthesis [8]. Moreover, the photosynthetic NUE (PNUE), which is dependent on the level of CO_2 saturation of Rubisco, is another important factor to consider when NUE is compared among different genotypes. At low N level, greater PNUE and NUE were found in C3 plants relative to C4 plants, whereas at high N level, the opposite is true [9]. Consequently, identification of the regulatory elements controlling the balance between N available to maintain photosynthesis and the reallocation of the remobilized N to sink organs such as developing young leaves is of major importance, particularly when N supply is restricted. Therefore, the complex regulators of N related to primary CO_2 assimilation, the photo-respiratory processes, and as storage pool need further investigation for optimizing NUE under low N level [10-12]. In addition, the recent finding that synthesis, turnover, and degradation of Rubisco are subjected to a complex interplay of regulations renews the concept of the importance of N use and recycling by the plants [13]. Attempts have also been made to identify some of the components responsible for the physiological control of the 'stay-green' phenotype particularly in relation to NUE. For example, in both sorghum and maize, delayed leaf senescence allowed photosynthetic activity to be prolonged, which had a positive effect on N uptake capacity of the plants [14-16].

In general, a low or zero nitrogen application causes nutritional imbalance. Plants can perceive the stress signals and transmit them to the cellular machinery to activate adaptive responses. The adaptation is generally completed by regulating gene expressions. Proteome dynamics under the stress conditions reflects the regulatory gene expressions. In the current study, in order to understand the NUE mechanism of rice, the largest food crop of the world, we adopted a systematic proteomic approach to investigate the nitrogen stress-responsive proteins in two rice cultivars differing in NUE.

2. Material and methods

2.1 *Plant materials and stress treatments*

Seeds of two rice cultivars, Chunyou 58 (high NUE) and Yongyou 6 (Low NUE), were germinated and grown hydroponically in nutrient solution containing 2.9 mM NH_4NO_3, 0.32 mM NaH_2PO_4, 1.0 mM K_2SO_4, 1.0 mM $CaCl_2$, 1.7 mM $MgSO_4 \cdot 7H_2O$, 9.1 μM $MnCl_2 \cdot 4H_2O$, 0.52 μM $(NH_4)_6Mo7O_{24} \cdot 4H_2O$, 18 μM H_3BO_3, 0.15 μM $ZnSO_4 \cdot 7H_2O$, 0.16 μM $CuSO_4 \cdot 5H_2O$, 36 μM $FeCl_3 \cdot 6H_2O$. The pH value of the solution was adjusted to 5.5 using 1 M HCl or NaOH solution as required [17]. Half concentration of the nutrient solution was applied for the first 3 days and then changed to full nutrient solution. At the emergence of the fourth leaf, the seedlings were transferred into either a nutrient solution without N supply as stress treatment or a nutrient solution with the normal N concentration as control. Nutrient solutions were renewed every four days. The upper expanded leaves were harvested after 12 hours, 3 days, and 7 days after the treatment, and kept frozen in liquid nitrogen and kept at -80 ℃.

2.2 *Sample preparation*

Leaf proteins were extracted by phenol extraction coupled with ammonium acetate precipitation [18]. Three separate extractions were conducted from three leaf samples of each treatment. Briefly, 1) 0.4 g frozen plant tissue with 30 mg PVPP was grinded into a fine powder using cold pestle and mortar. 2) Suspend the powder in cold phenol extraction buffer (0.7 M sucrose; 0.1 M KCl; 0.5 MTris-HCl, pH 7.5 and 50 mM EDTA, 1% w/v DTT, pH 7.5; complete protease inhibitor cocktail (Roche Applied Science)), then add an equal volume of phenol saturated with Tris-HCl, pH 7.5, and after 30 min shake the mixture at 4℃. 3) Centrifuge at 5,000 g for 30 min at 4℃, then collect the upper phenolic phase; discard the lower aqueous phase. 4) Add extraction buffer to the collected phenolic phase; repeat steps 5-6 and then repeat Step 5 again. 5) Add 5 volumes of cold 0.1 M ammonium acetate in methanol to the collected phenol phase; stored at –20℃ overnight. 6) Centrifuge the sample for 30 min at 5000 g at 4℃ and carefully remove the supernatant with a pipette and discard. 7) Add 2 volumes (based on the volume of the last collected phenolic phase) of ice-cold methanol to wash

the pellet, centrifuge the sample for 10 min at 5,000 g at 4℃; repeat step 7 two more times to remove ammonium acetate and phenol, lipids and pigments and repeat step 9 twice again using acetone instead of methanol to replace. 8) Dry the pellet gently in a fume hood, and store the clear supernatants in aliquots at -80℃ until analyzed. The protein concentrations were measured by a Bradford assay using bovine serum albumin as standard (Bio-Rad, Hercules, CA, USA).

2.3 Two-dimensional electrophoresis

Two-DE was performed according to the manual obtained from GE Healthcare Life Sciences (Little Chalfont, United Kingdom). Extracellular protein preparation containing 150 μg protein was separated by 2-DE using 24 cm immobilized pH gradient strips pH4-7 (GE Healthcare, Milwaukee, WI, USA). Briefly, sample was diluted with rehydration solution (8 M Urea, 2 M Thiourea, 4% w/v CHAPS, 20 mM w/v DTT, 0.5% v/v IPG buffer pH 4–7, 0.002% w/v bromophenol blue) to 0.5-1 mg protein per 100 μl. Immobiline DryStrip gels (pH 4-7, 24 cm; GE Healthcare Life Sciences) were then rehydrated with 450 mL of mixture solutions in 17 cm strip holders and electrofocused with the GE Healthcare Life Sciences IPGphor. Initially, all protein extracts were subjected to 2-DE. Among the three biological replicates of each leaf-treatment sample, the one with best 2-DE quality were chosen to run two times more 2-DE. About 100 mg of protein were loaded using in-gel rehydration. The focusing protocol was as follows: 50 mA per strip at 20 ℃; (i) rehydration with 30 V for 12 h;(ii) 500 V for 1 h (step and hold);(iii) 1000 V for 1 h (step and hold); and (iv) 8000 V for 10 h (step and hold) was applied until the total Vh reached 100 kVh. After IEF, the strips were equilibrated twice with gentle shaking for 15 min in SDS equilibration buffer.

The first step was performed in a equilibration solution containing 6 M urea, 30% w/v glycerol, 2% w/v SDS, 1% w/v DTT, 50 mM Tris-HCl buffer, pH 8.8 and 0.002% w/v bromophenol blue. The second step was performed in a solution modified by the replacement of DTT with 2.5% w/v iodoacetamide. When the equilibration was finished, the strips were loaded onto vertical SDS PAGE (12.5% T constant). The second dimension SDS electrophoresis was run using an Ettan DALTsix electrophoresis Unit (Amersham Biosciences). A denaturing solution (0.5% Agarose in running buffer) was loaded onto the gel strips and electrophoresis was performed in a Laemmli running buffer (25 mM Tris–HCl pH 8.3, 192 mM glycine, 0.1% SDS). The gels were run at 2-2.5 W per gel for the first 40 min and followed by 17 W per gel for 6 h until the dye front reached the bottom of the gel. For quantitative analysis of protein abundance profiles, gels were stained by silver-staining according to the manufacturer's instructions (GE Healthcare, Milwaukee, WI, USA). The stained gels were scanned in an ImageScanner (PowerLook1100 scanner, UMAX) and were analyzed with ImageMaster 2D Elite software. The three technical replicates of each biological sample were pooled and averaged.

2.4 In-gel digestion and MALDI-TOF MS analysis

Spot detection was realized without spot editing. The spots were quantified using the % volume criterion. Only those with significant and reproducible changes were considered to be differentially accumulated proteins. Protein spots were excised from the Silver-stained gels and transferred into 0.2 mL Eppendorf tubes. Each spot was washed twice in milli-Q water (Millipore), destained by washing with 50% MeOH/ 50mM NH_4HCO_3 for 30 min. The gels were then washed twice in milli-Q water, dehydrated by addition of ACN (acetonitrile, Fisher A/0626/17), and dried in a SpeedVac (Thermo Savant, Holbrook, NY, USA) for 30 min. Subsequently, the gel in each tube was rehydrated in 3 μl of proteomics grade trypsin (Sigma) solution (20 ng/mL 40 mM NH_4HCO_3 in 9% ACN) and incubated at 37℃ for 16 h. Peptides were extracted twice by adding 40 μl of solution containing 50% can and 5% TFA (trifluoroacetic acid, GE HealthCare). The extracted solutions were concentrated to 5-10 μl in a lyophilizer (Virtis, Gardiner, NY, USA). Peptide mixtures were mixed with an equal volume of 10 mg/mL CHCA (Sigma) saturated with 50% ACN in 0.1% TFA (Sigma) and analyzed with a Voyager-DE STR MALDI-TOF-TOF mass spectrometer (ABI4700 System, USA) using a delayed ion extraction and ion mirror reflector (Applied Biosystems, Foster city, CA, USA). MS analysis was conducted with a MALDI-TOF/TOF mass spectrometer 4700 Proteomics Analyzer (Applied Biosystems, Framingham, MA, USA). Data were analyzed using GPS Explorer software (Applied Biosystem) and MASCOT software (Matrix Science, London, UK). Parameters were set to Variable Modification - Oxidation, 1 Allowed Missed Cleavage. NCBInr and *Oryza sativa* (rice) was selected as the database and taxonomy, respectively.

3. Results and Discussion

Physiological response to nitrogen stress

Four-leaf age seedlings were exposed to N-free nutrient solution and sampled at different times. Obvious nitrogen stress symptoms were observed in the seedlings, such as yellow leaves and impaired tillers. Highly significant difference in nitrogen content and accumulation were also found between the N stress treatment and the control (Table 1). In normal condition with sufficient N supply, cultivar Yongyou 6 had higher dry weight than Chunyou 58, which was consistent with the difference in nitrogen accumulation between the two cultivars. However, when the seedlings were exposed to N stress, Yongyou 6 showed higher loss of dry weight than Chunyou 58. The two cultivars both showed significant decline of nitrogen content and accumulation under N stress relative to the normal condition.

2-DE analysis of leaf proteins in nitrogen stressed rice

Total proteins in the fully-expanded leaves were extracted and separated by 2-DE using pH 4–7 IPG strips in IEF. More than 1,000 protein spots were reproducibly detected on gels by ImageMaster 2D Elite software. Spots with biological

Table 1. Shoot dry weigh and N content and accumulation of the rice cultivars under the different N treatments.

Cultivar	Treatment	Shoot dry weight (g/pot)			N content (%)			Shoot N accumulation(mg/pot)		
		12h	3d	7d	12h	3d	7d	12h	3d	7d
Chunyou 58	0-N	0.14a	0.21a	0.42b	6.24a	4.22b	2.04b	8.7a	8.9a	8.6b
	Control	0.15a	0.22a	0.53a	6.34a	5.37a	4.54a	9.2a	11.8a	24.0a
Yongyou 6	0-N	0.28a	0.44b	0.58b	6.46a	4.21b	3.27b	18.1a	18.5b	18.9b
	Control	0.30a	0.65a	0.82a	6.37a	5.14a	4.00a	19.1a	33.4a	32.8a

significance (ratio > 1.3) between the two treatments are showed in Figure 1. In order to investigate changes in protein accumulation profiles between the control and N-stressed rice plants, the ratio of differentially accumulated proteins between N stress and the normal treatments was calculated, and the proteins with the ratio of over 1.3 were further examined (Fig. 2). Apparently, there were more proteins, which showed significant and reproducible changes in Yongyou 6 than in Chunyou 58. In addition, the two cultivars differed greatly in the number of differential proteins (up or down regulation) over the time of treatment. Chunyou 58 reached the maximum differential proteins in 3 d after the treatment of N stress, while Yongyou 6 did not show the obvious difference over the time of treatment.

There were 31 protein spots in the two cultivars that showed reproducible changes during the treatment, and were selected for MALDI-TOF MS analysis. Among them, 2 and 11 were down-regulated spots for Chunyou 58 (C-D1, C-D2) (Fig. 2A) and Yongyou 6 (Y-D1-Y-D10) (Fig. 2B), respectively; and 8 and 16 were up-regulated spots for Chunyou 58 (C-U1–C-U6) (Fig. 2A) and Yongyou 6 (Y-U1-Y-U11) (Fig. 2B), respectively. The abundance ratios, i.e. the percentage volumes in treated samples over the percentage volumes in control samples, in different stages are shown in Figure 3. Qualitative changes of spots have been found. For example, Y-D6 was visible in all stages of the N stress treatment but invisible in control samples, suggesting that it was induced under N stress treatment. Some differentially accumulated proteins showed quantitative changes in a time-dependent manner. For instance, C-U5, Y-U5, Y-U4 and Y-U16 showed smaller difference between the treated and control samples in early stages of N stress, such as at 12 h. Their abundance ratios were greater at 3 d or 7 d (Fig. 3), indicating that the syntheses of the proteins in the treated sample were enhanced. In addition, spot Y-U4 was observed with a dramatic increase in the abundance in the treated sample at 3 d, while spot Y-D7 decreased dramatically and almost disappeared in the treated sample at 7 d.

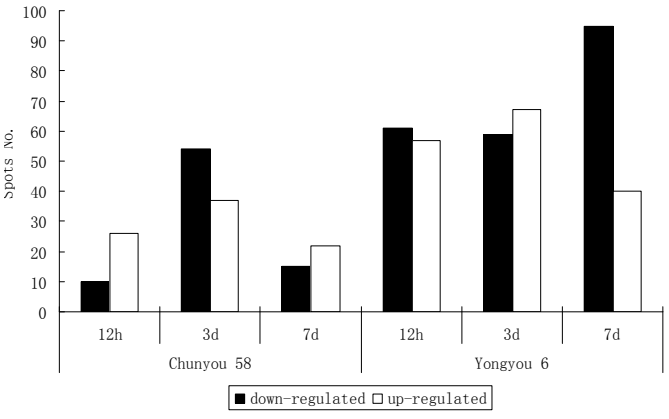


Figure 1. Number of spots whose abundance ratio of the differentially accumulated proteins were over 1.3 after N stress treatment. The percentage volume was considered as the abundance of each spot. The abundance ratio of each spot was calculated by percentage volume in treated samples/ percentage volume in control samples as up-regulated spots (□), while the ones was calculated by percentage volume in control sample/ percentage volume in treated sample as down-regulated spots (▪).

N stress responsive proteins identified by MS

A total of 37 differentially accumulated protein spots were analyzed and identified by MALDI-TOF/ TOF MS with high probability (Table 2). "Spots view" of 15 protein spots of time-dependent changes was shown in Figure 4 as examples. Four identified proteins were found in both varieties in all times during the stress treatment (Table 2). Spots C-U1 and Y-U6 were identified as the same protein, ribulose-1, 5-bisphosphate carboxylase/oxygenase activase. However, they were located at different positions on the gels, with different Mr and pI (Fig. 2 A and B), indicating that they might be isoforms of ribulose-1, 5-bisphosphate carboxylase/ oxygenase activase. It can be assumed that the enzyme is up-regulated under stress since its expression is enhanced with decreased RuBisCo abundance which will reduce photosynthesis. Spots C-D1 and Y-D10 were identified as rubisco large subunit with similar Mr and pI. Spots C-U8 and Y-U15 were identified as H protein subunit of glycine decarboxylase 3'-partial. Spots C-D2 and Y-D5 were identified as putative transposase.

Five proteins were involved in photosynthetic metabolism, including ribulose-1,5-bisphosphate carboxylase/oxygenase activase (C-U1 and Y-U6), type II light-harvesting chlorophyll a/b-binding protein (C-U4), carbonic anhydrases (C-U5), rubisco large subunit (C-D1 and Y-D10), 23kDa polypeptide of photosystem II (Y-U9), dTDP-glucose 4-6-dehydratase-like protein (C-U7) and H protein subunit of glycine decarboxylase 3'-partia (C-U8 and Y-U15). Six proteins were the stressor response to N stress i.e. DegP2(Y-D6), harpin binding proteins(Y-D11), Heat shock-related proteins

(Y-U2), glutathione S-transferase GSTF14(Y-U4), Fibrillin-like protein(Y-U6) in Yongyou 6, and Glyceraldehyde-3-phosphate dehydrogenase (C-U3) in Chunyou 58.

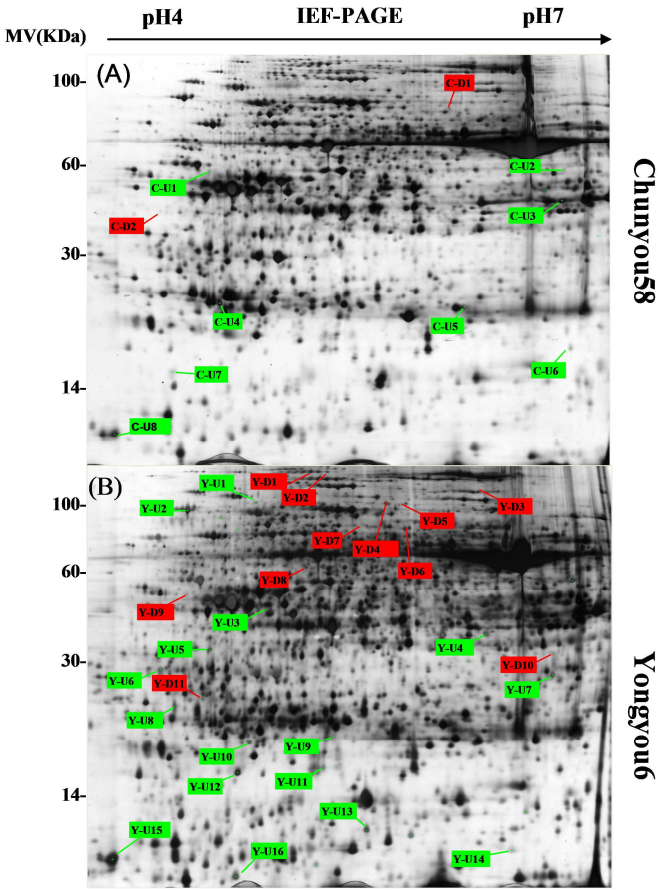


Figure 2. Representative 2-DE maps of rice leaf proteins. Differentially accumulated protein spots (Ratio>1.3) which appeared in all time are selected for MALDI-TOF MS analysis and indicated by label in the map. Two and twelve down-regulated spots for Chunyou58 (C-D1, C-D2) and Yongyou6 (Y-D1-Y-D10) are indicated by red marker. Ten and twenty four up-regulated spots for Chunyou58 (C-U1–C-U6) and Yongyou6 (Y-U1-Y-U11) are indicated on the map by green marker

Photosynthesis and photorespiration

Rubisco activase (RCA, spots C-U1 and Y-U6) is the key enzyme for the rapid formation of the critical carbamate in the active site of RuBisCo. It is modulated either by reaction with CO_2 and Mg^{2+} to carbanylate a lysine residue in the catalytic site, or by the binding of inhibitors within the catalytic site [19]. A variable number of RCA genes have been reported in different plant species. In monocotyledonous plants, two genes have been detected [20]. Two mature RCA polypeptides, with molecular mass ranging between 41 kDa and 47 kDa are present in most plants [21]. Increased 43 Kda/41 KDa was found in low light intensity whereas the decreased one was found in water-stress [20]. Our result showed that RCA (C-U1 and Y-U6) was strongly up-regulated by N stress in most expanded leaves of both rice cultivars (Table 2). N stress can reduce the photosynthesis either by impairing activation state of RuBisCo, which is often attributed to the loss of RCA activity or by reducing the abundance of RuBisCo proteins. The reduced RuBisCo was proved by a dramatic down- regulation of RuBisCo large subunit (RLS) (C-D1 and Y-D10) in N stress samples (Table 2). These results suggest that the RCA was over-expressed as a feedback mechanism for decrease of RCA activity or RuBisCo content in both rice cultivars. However, different additional protective strategy was found between Chunyou 58 and Yongyou 6. Type II light-harvesting chlorophyll a/b-binding protein (LHCP) (C-U4) and carbonic anhydrases (CA) (C-U5) were found in the N treated Chunyou 58 and 23 kDa polypeptide of photosystem II (PsbP-PSII) (Y-U9) was strongly up-regulated in treated samples of Yongyou 6. LHCP is an approximately 25,000-D thylakoid membrane protein, which captures and transmits the energy from the sunlight into biomass [22]. The increased LHCP was found under N stress in this experiment. With the reduced photosynthetic efficiency resulting from N stress, the up-regulated LHCP in N stressed sample might be a compensation for the low photosynthetic efficiency in order to capture and transform more energy to produce the carbohydrate for plant growth. AC forms a family of enzymes, which catalyze rapid conversion of carbon dioxide to bicarbonate and protons. In plants, AC may increase CO_2 concentration within chloroplasts in order to enhance carboxylation rate of RuBisCO [23]. Rengel (1995) found that higher photosynthetic rate under Zn deficiency was related to higher CO_2 availability due to higher CA activity in some wheat genotypes [24]. It may be assumed that the less reduction of biomass in N-stressed Chunyou 58 comparing to Yongyou 6 might be attributed to higher CA activity. PsbP-PSII is one of subunits of the oxygen-evolving complex (OEC) of PSII. The 23-kDa subunit allows PSII to evolve oxygen under both Ca^{2+} and Cl^{2-} limiting conditions, suggesting that it acts as a concentrator of these ions [10]. The dramatic up-regulation of PsbP-PSII was found in N-stressed samples of Yongyou 6, which may act as a compensation for the decreased photosynthesis induced by N stress. The difference in regulated metabolisms between the two rice cultivars may be one of the major causes that lead to more biomass reduction for Yongyou 6 than for Chunyou 58.

A marked increase in H protein subunit of glycine decarboxylase 3'-partial was found in N-stressed rice plants (H-GDC) (Chunyou 58, C-U8; Yongyou 6, Y-U15). H-protein is the lipoyl-protein component of the glycine decarboxylase complex (GDC), which oxidizes glycine to support photorespiration [25]. GDC consists of four proteins, including H-protein that helps to receive the released CO_2. It was reported that the expression of H-protein gene in leaf was stimulated by light [26]. In this experiment, more H-protein was found in N-stressed samples, indicating that photorespiration might be enhanced when the plants are exposed to N stress.

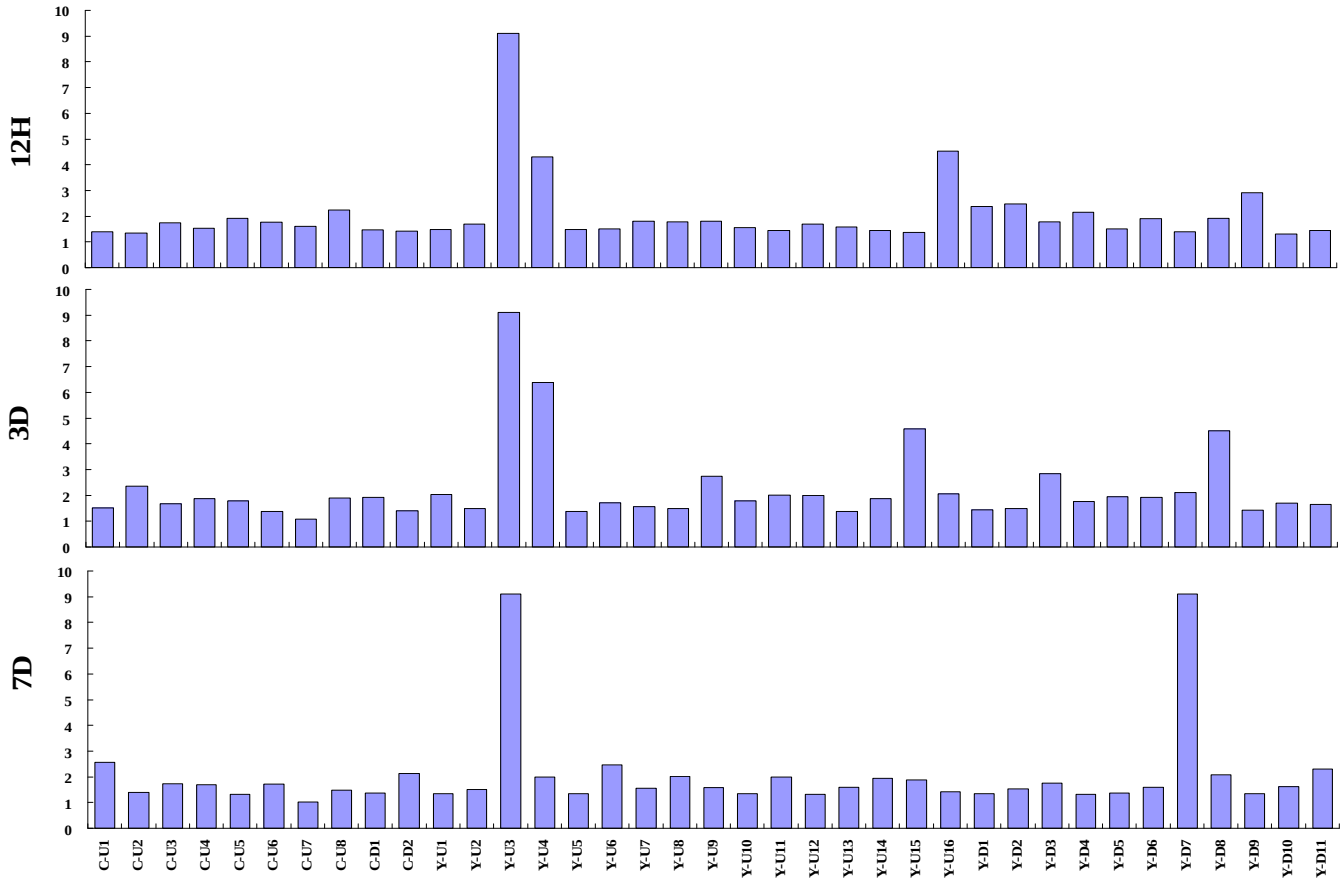


Figure 3. Abundance ratios of the differentially accumulated proteins after 12 h, 3 d and 7 d of N stress treatment. The percentage volume was considered as the abundance of each spot. The abundance ratio of each spot was calculated by percentage volume in treated samples / percentage volume in control samples. The up-regulated proteins include C-U1 to C-U6 (A, D, G), and Y-U1 to Y-U11 (B, E, H); the down-regulated proteins are C-D1 to C-D2 (A, D, G), and Y-D1 to Y-D10 (C, F, I). Spots with * means either the abundance ratio of differentially accumulated protein was over 10,000 or the protein was absent in the treated or control sample.

N stress- induced proteins

Many N stress-related proteins were identified in this study. In general, Yongyou 6 was more sensitive to N stress than Chunyou 58. There were 5 N stress-induced proteins in Yongyou 6, including two down-regulated ones: DegP2(Y-D6) and harpin binding proteins(Y-D11), and three up-regulated ones: heat shock-related proteins(Y-U2), glutathione S-transferase GSTF14(Y-U4), and Fibrillin-like protein(Y-U6). On the contrast, only one stress-related protein was found in Chunyou 58, i.e. glyceraldehyde-3-phosphate dehydrogenase (C-U3). DegP2 is a member of a large family of related Deg/Htr serine proteases found in most organisms, including bacteria [27], humans [28] and plants [29]. Bacterial DegP/HtrA protease has been implicated in tolerance to various stresses, including oxidation, salinity, pH and heat [28]. The current results showed that this protein was reduced under N stress (Table 2), suggesting that the effect of DegP/HtrA protease on enhancing plant tolerance under the stress condition relies on the nitrogen nutrition. Fibrillin-like protein is a glycoprotein, which is essential for the formation of elastic fibers [30]. As lipid-binding proteins of plastids, fibrillin are induced under abiotic stress conditions. Yang (2006) reported that ABA treatment increased fibrillin accumulation, thus enhancing the tolerance of photosystem II to light stress-triggered photo-inhibition in *Arabidopsis* [31]. In this study, fibrillin–like protein were increased in N stressed plants of Yongyou 6, indicating that as a feedback mechanism for N deficiency, the efficiency of photosynthesis was improved by inducing more fibrillin proteins to protect the photosystem II. The harpin protein group, which is first found and identified by Wei et al. (1992) in *Erwinia amylovora* [32], may elicit multiple plant responses, causing beneficial effects on crop improvement [33]. The current results indicated the possible defense mechanism of rice plants in response to N stress by inducing harpin proteins, thus enhancing photosynthesis and nitrogen uptake. Heat shock-related proteins (HSP) are a class of functionally related proteins, whose expression is increased when cells are exposed to elevated temperatures or other stresses [34]. The function of glutathione S-transferase GSTF14 protects cells from injury by a wide range of stresses in plants [35]. A significant up-regulation of heat shock related proteins and GSTF14 was found in N stressed Yongyou 6 (Table 2).

Table 2. Differentially accumulated proteins identified by MS.

MS SPOT	Rank Protein Name	Accession No.	Protein Score	Protein MW	Protein PI	Ratio		
						12h	3d	7d
C-U1	ribulose-1,5-bisphosphate carboxylase/oxygenase activase [Oryza sativa]	gi|115486823	218	48127.9	5.85	1.40	1.52	2.55
C-U2	putative gypsy-type retrotransposon [Oryza sativa (japonica cultivar-group)]	gi|18071410	39	165471.2	9.53	1.36	2.36	1.40
C-U3	glyceraldehyde-3-phosphate dehydrogenase [Oryza sativa]	gi|2331137	173	36707.0	9.55	1.74	1.66	1.75
C-U4	type II light-harvesting chlorophyll a/b-binding protein [Oryza sativa Japonica Group]	gi|218174	54	28566.4	5.61	1.54	1.88	1.68
C-U5	carbonic anhydrase 3 [Oryza sativa]	gi|5917783	134	29585.0	8.41	1.92	1.79	1.34
C-U6	putative metalloproteinase [Oryza sativa (japonica cultivar-group)]	gi|14165330	42	32250.9	6.36	1.75	1.38	1.72
C-U7	dTDP-glucose 4-6-dehydratase-like protein [Oryza sativa]	gi|18201659	40	26091.4	7.1	1.60	1.08	1.03
C-U8	H protein subunit of glycine decarboxylase 3'-partial [Oryza sativa (japonica cultivar-group)]	gi|10257441	67	7156.6	8.98	22.25	11.89	1.50
C-D1	rubisco large subunit	gi|476752	79	45614.8	8.43	1.46	1.91	1.38
C-D2	putative transposase [Oryza sativa (japonica cultivar-group)]	gi|34015353	49		6.56	1.42	1.39	2.13
Y-U1	ribulose-1,5-bisphosphate carboxylase activase [Oryza sativa]	gi|13569643	273	21737.8	4.78	1.49	2.04	1.35
Y-U2	heat shock-related protein [Oryza sativa (japonica cultivar-group)]	gi|29367425	404	45014.5	5.02	1.69	1.48	1.52
Y-U3	Os06g0176700 [Oryza sativa (japonica cultivar-group)]	gi|115466716	203	40022.5	5.16	1000000	1000000	1000000
Y-U4	glutathione S-transferase GSTF14 [Oryza sativa(japonica cultivar-group)]	gi|46276327	514	30766.5	7.77	4.52	6.38	2.00
Y-U5	putative protein kinase ADK1 [Oryza sativa Japonica Group]	gi|52077492	41	26201.6	9.55	1.49538	1.37301	1.34866
Y-U6	fibrillin-like protein [Oryza sativa (japonica cultivar-group)]	gi|29367475	510	33923.7	5.04	1.51	1.72	2.47
Y-U7	Putative wall-associated protein kinase [Oryza sativa (japonica cultivar-group)]	gi|14029040	40	53505.1	6.01	1.8159	1.55712	1.55934
Y-U8	oryzain gamma precursor [Oryza sativa Japonica Group]	gi|218185	51	39692.5	7.07	1.79	1.48	2.01
Y-U9	23kDa polypeptide of photosystem II [Oryza sativa]	gi|2570499	271	27173.9	9.06	1.82	2.74	t
Y-U10	Os08g0455800 [Oryza sativa (japonica cultivar-group)]	gi|115476734	240	21683.7	5.15	1.56	1.79	1.34
Y-U11	hypothetical protein [Oryza sativa Japonica Group]	gi|42407348	54	7766.8	10.96	1.44	2.02	2.00
Y-U12	Os08g0478200 [Oryza sativa (japonica cultivar-group)]	gi|115476908	371	19712.9	5.19	1.70	1.98	1.33
Y-U13	Os10g0471300 [Oryza sativa (japonica cultivar-group)]	gi|115482468	161	18653.4	5.61	1.57	1.38	1.61
Y-U14	hypothetical protein LOC_Os03g43310 [Oryza sativa (japonica cultivar-group)]	gi|53370666	39	20432.4	10.86	1.45	1.87	1.94
Y-U15	H protein subunit of glycine decarboxylase 3'-partial [Oryza sativa (japonica cultivar-	gi|10257441	60	7156.6	8.98	1.37	4.58	1.88

	group)]							
Y-U16	Os06g0705100 [Oryza sativa (japonica cultivar-group)]	gi|115469830	298	24997.9	8.74	1000000	2.06	1.42
Y-D1	putative chloroplast inner envelope protein [Oryza sativa (japonica cultivar-group)]	gi|10140720	812	108209.8	5.37	2.38	1.43	1.35
Y-D2	putative SecA [Oryza sativa Japonica Group]	gi|52075758	410	114899.0	5.78	2.47	1.49	1.54
Y-D3	Os03g0401300 [Oryza sativa (japonica cultivar-group)]	gi|115453437	72	93362.3	5.94	1.79	2.84	1.77
Y-D4	Os02g0285800 [Oryza sativa (japonica cultivar-group)]	gi|115445587	360	74035.2	7.08	2.16	1.76	1.32
Y-D5	putative transposase [Oryza sativa (japonica cultivar-group)]	gi|34015353	45	66527.4	6.56	1.51	1.94	1.37
Y-D6	putative DegP2 protease [Oryza sativa (japonica cultivar-group)]	gi|51038169	37	65771.4	5.73	1.94	1.93	1.61
Y-D7	Os06g0562600 [Oryza sativa (japonica cultivar-group)]	gi|115468554	268	59720.8	5.71	1.39	2.11	1000000
Y-D8	eukaryotic initiation factor 4A [Oryza sativa Japonica Group]	gi|303844	207	47393.1	5.43	1.91	4.52	2.08
Y-D9	hypothetical protein [Oryza sativa (japonica cultivar-group)]	gi|13236651	142	45168.2	5.27	2.90	1.43	1.35
Y-D10	rubisco large subunit	gi|476752	157	45614.8	8.43	1.31	1.70	1.63
Y-D11	harpin binding protein 1 [Oryza sativa (indica cultivar-group)]	gi|38679325	78	28457.0	8.92	1.43	1.65	2.30

Glyceraldehyde 3-phosphate dehydrogenase (C-U3) (GAPDH, EC 1.2.1.1) was up-regulated in N stressed Chunyou 58. GAPDH plays important roles in various cellular processes. It is a central glycolytic protein with pivotal role in energy production, and is also an abundant and crucial enzyme in glycolysis and gluconeogenesis in most plants [36]. Moreover, GAPDH is a protein with multi-function, involving in the translational control of gene expression [37]. For the last decade, there were many reports that GAPDH works as a stressor associated with oxidative stress in cells that undergo apoptosis [38]. It may be suggested that over-expression of GAPDH in N stressed Chunyou 58 acts not only as an oxidative signal to N deficiency, but also an energy production through its glycolytic function.

Membrane transporter

Putative chloroplast inner envelop protein (Y-D1) and SecA protein (Y-D2) were down-regulated in N stressed Yongyou 6. Chloroplast inner envelop is highly specialized with transport proteins, which involved in the movement of ions, small molecules, or macromolecules. SecA proteins were found in the thylakoid membrane as well as the cytoplasmic membrane, and they involved in protein translocation across the thylakoid membrane [39]. The current results showed that the translocation across the chloroplast or thylakoid membrane was inhibited in Yongyu 6 under N stress.

Different protective strategies under N stress between two cultivars

Resolving the cause and effect relationship in plants subject to a nitrogen limitation is difficult because nitrogen stress initiates a series of complex physiological responses varied over the time and the stress degree. Many of the physiological metabolisms were directly or indirectly involved with the stress effect. For example, the net photosynthesis rate in most extended leaves acclimate to N stress with highly up-regulated RCA in two cultivars as well as LHCP / CA and PsbP-PSII in Chunyou 58 and Yongyou 6, respectively. The similar results were also reported by De Groot et al. (2003) that photosynthetic light-harvesting and electron-transport activity acclimate to nitrogen stress so that the internal relationships between electron transport by photosystems I and II do not change; the linear relationship between PSII, and PSI was not affected [40]. In protein profile, RCA could play a role of a chaperone, either in helping target the thylakoid membrane or in protection of translation machinery related to thylakoid against abiotic stress [41]. Thus the up-regulated RCA could protect the photosynthesis machinery under N stress. Despite of the similar protection from the RCA, different protective strategies were drown out from cultivars differing in NUE. LHCP / CA were up regulated in Chunyou 58 resulting in the slightly biomass decrease subjected to a nitrogen limitation. Only PsbP-PSII was consistently up regulated for carbon production metabolism over the stress period in Yongyou 6. This may be considered as the major cause for the biomass differentiation between the cultivars.

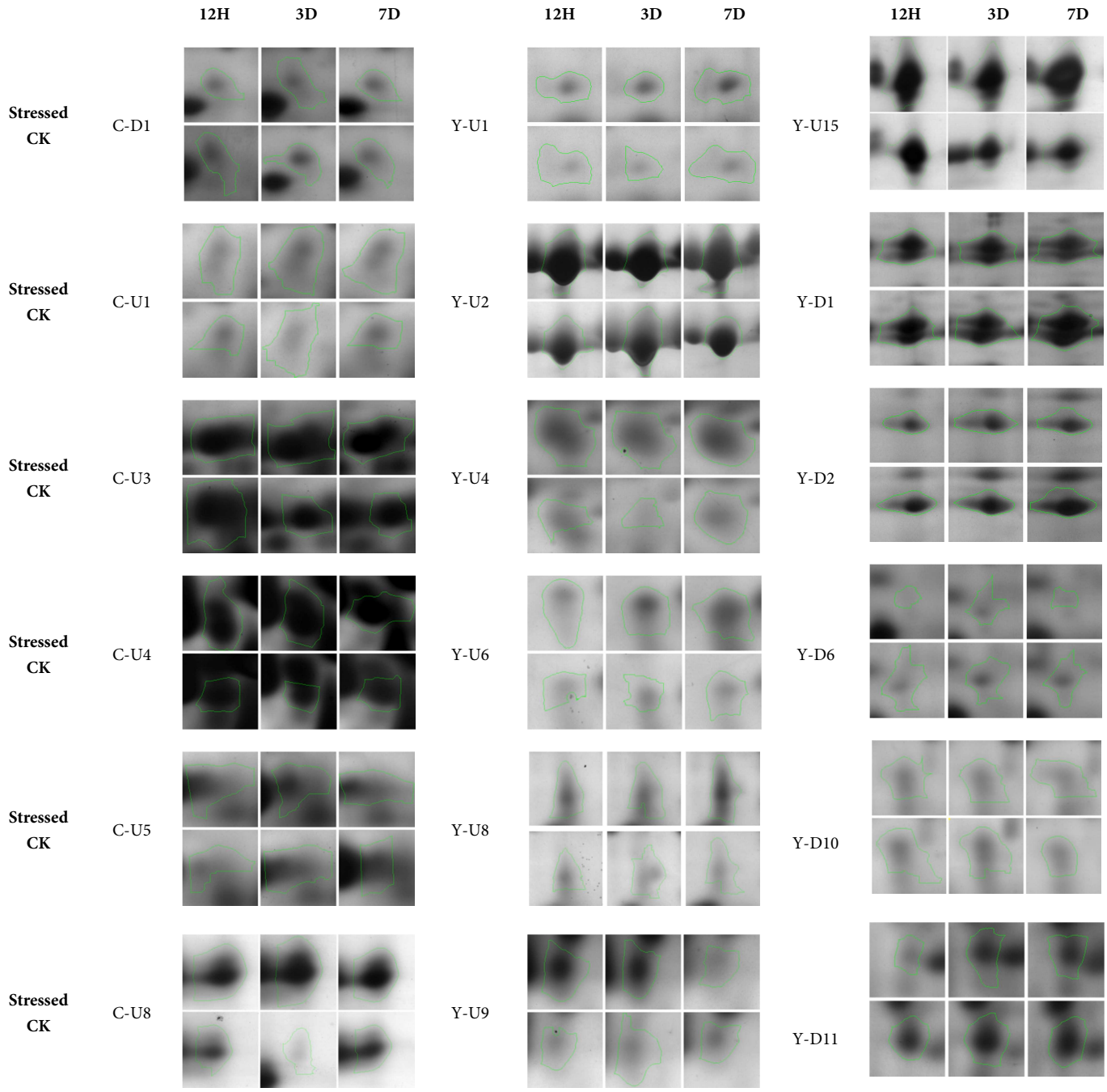


Figure 4. Time-dependent changes of the 15 of 48 differentially accumulated proteins. Proteins in leaves were extracted from both control and stressed samples after 12h, 3d and 7d treatment and separated by 2-DE.

4. Concluding remarks

A systematic proteomic analysis of the leaf proteins in N stressed rice was carried out in this study. Of the six protein spots involved in photosynthesis and photorespiration, three were identified in the two rice cultivars at all times during the treatment. These are: Rubisco activase, RuBisCo large subunit, and H protein subunit of glycine decarboxylase 3'-partial. Six stress-induced proteins were identified, including DegP2, harpin binding proteins, heat shock-related proteins, glutathione S-transferase GSTF14, fibrillin-like protein and glyceraldehyde-3-phosphate dehydrogenase. Two proteins, i.e. putative chloroplast inner envelop protein and SecA protein are related to membrane translocation, Moreover, two novel proteins, harpin binding protein and oryzains gamma precursor, were found in the rice leaves under N stress. These results provide useful information for further investigation of their functions using genetic or genomic approaches.

Acknowledgements

We thank Bureau of Science and Technology of Zhejiang, China for its financial support to this research project and Dr. Younan Ouyang of Chinese Rice Research Institute for his kind help in nitrogen measurement.

References

1. Novoa, R. and Loomis, R. S., Plant and Soil 1981, 58, 177-204.
2. Tilman, D., Proceedings of the National Academy of Sciences of the United States of America 1999, 96, 5995-6000.
3. Raun, W. R. and Johnson, G. V., Agronomy Journal 1999, 91, 357-363.
4. Frink, C. R., Waggoner, P. E. and Ausubel, J. H., Proceedings of the National Academy of Sciences of the United States of America 1999, 96, 1175-1180.
5. Mae, T., Plant and Soil 1997, 196, 201-210.
6. Peng, S., Cassman, K. G., Virmani, S. S., Sheehy, J. and Khush, G. S., Crop Science 1999, 39, 1552-1559.
7. Dalling, M. J., 1985.
8. Gastal, F. and Nelson, C. J., Plant Physiology 1994, 105, 191-197.
9. Sage, R. F., Pearcy, R. W. and Seemann, J. R., Plant Physiology 1987, 85, 355-359.
10. Murata, M., Richardson, J. S. and Sussman, J. L., Proceedings of the National Academy of Sciences of the United States of America 1985, 82, 3073-3077.
11. Esquivel, M. G., Ferreira, R. B. and Teixeira, A. R., Plant Science 2000, 153, 15-23.
12. Lawlor, D. W., Journal of Experimental Botany 2002, 53, 773-787.
13. Hirel, B. and Gallais, A., New Phytologist 2006, 169, 445-448.
14. Ma, B. L. and Dwyer, L. M., Plant and Soil 1998, 199, 283-291.
15. Rajcan, I. and Tollenaar, M., Field Crops Research 1999, 60, 245-253.
16. Rajcan, I. and Tollenaar, M., Field Crops Research 1999, 60, 255-265.
17. Yoshida, S., Forno, D. A., Cock, J. H. and Gomez, K. A., Laboratory manual for physiological studies of rice 1976, 3, 61–66.
18. Isaacson, T., Damasceno, C. M. B., Saravanan, R. S., He, Y., et al., Nature Protocols 2006, 1, 769-774.
19. Portis, A. R., Photosynthesis Research 2003, 75, 11-27.
20. Rundle, S. J. and Zielinski, R. E., Journal of Biological Chemistry 1991, 266, 14802-14807.
21. Salvucci, M. E., Werneke, J. M., Ogren, W. L. and Portis, A. R., Plant Physiology 1987, 84, 930-936.
22. Keegstra, K., Olsen, L. J. and Theg, S. M., Annual Review of Plant Physiology and Plant Molecular Biology 1989, 40, 471-501.
23. Badger, M. R. and Price, G. D., Annual Review of Plant Physiology and Plant Molecular Biology 1994, 45, 369-392.
24. Rengel, Z., Journal of Plant Physiology 1995, 147, 251-256.
25. Kikuchi, G., Molecular and Cellular Biochemistry 1973, 1, 169-187.
26. Macherel, D., Bourguignon, J. and Douce, R., Biochemical Journal 1992, 286, 627-630.
27. Gottesman, S., Annual review of genetics 1996, 30, 465-506.
28. Spiess, C., Beil, A. and Ehrmann, M., Cell 1999, 97, 339-347.
29. Itzhaki, H., Naveh, L., Lindahl, M., Cook, M. and Adam, Z., Journal of Biological Chemistry 1998, 273, 7094-7098.
30. Kielty, C. M., Baldock, C., Lee, D., Rock, M. J., et al., Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 2002, 357, 207-217.
31. Yang, Y., Sulpice, R., Himmelbach, A., Meinhard, M., et al., Proceedings of the National Academy of Sciences of the United States of America 2006, 103, 6061-6066.
32. Wei, Z. M., Laby, R. J., Zumoff, C. H., Bauer, D. W., et al., Science 1992, 257, 85-88.
33. Alfano, J. R. and Collmer, A., Annual Review of Phytopathology 2004, 42, 385-414.
34. De Maio, A., Shock 1999, 11, 1-12.
35. Marrs, K. A., Annual Review of Plant Physiology and Plant Molecular Biology 1996, 47, 127-158.
36. Sirover, M. A., Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology 1999, 1432, 159-184.
37. Mezquita, J., Pau, M. and Mezquita, C., Journal of Cellular Biochemistry 1998, 71, 127-139.
38. Saunders, M., Dische, S., Barrett, A., Harvey, A., et al., Lancet 1997, 350, 161-165.
39. Nakai, M., Goto, A., Nohara, T., Sugita, D. and Endo, T., Journal of Biological Chemistry 1994, 269, 31338-31341.
40. De Groot, C. C., Van Den Boogaard, R., Marcelis, L. F., Harbinson, J. and Lambers, H., J Exp Bot 2003, 54, 1957-1967.
41. Law, R. D., Crafts-Brandner, S. J. and Salvucci, M. E., Planta 2001, 214, 117-125.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.33

Adaptive-BLAST: A User-defined Platform for the Study of Proteins

Yoojin Hong[1,2,3], Sree V Chintapalli[2,3], Gaurav Bhardwaj[2,3], Zhenhai Zhang[2,3], Randen L. Patterson*[,2,3], Damian B. van Rossum*[2,3]

[1]Department of Computer Science and Engineering, The Pennsylvania State University, USA; [2]Center for Computational Proteomics, The Pennsylvania State University, USA; [3] The Huck Institute of Life Sciences, The Pennsylvania State University, USA.

Received: 29 August 2010 **Accepted:** 11 September 2010 **Available Online**: 1 October 2010

ABSTRACT

Profile-based protein-sequence analysis algorithms comprise some of the most powerful and user-friendly methods for exploring protein sequences to determine their structure, function, and/or evolution (1-4). PSI-BLAST (5, 6) and rps-BLAST (7) are two of the most popular profile-based algorithms (~1,120 references to date), and have exceptional utility in the identification of homology between proteins, particularly for biological scientists who do not specialize in computational approaches. However, when the performance of these algorithms is compared to other methods [e.g. support-vector machine learning (SVM) (8), hidden-Markov models (HMMs) (9)], they often underperform in identifying the aforementioned protein properties (3, 9-11). We have previously demonstrated that the utility of BLAST algorithms can be significantly improved by: (i) adaptations to the profile libraries employed, (ii) adjustments to output formats, and (iii) alterations to BLAST algorithm itself (4, 6, 12-14). We present here Adaptive-BLAST (Ada-BLAST), which provides a simple user-defined platform for measuring and analyzing primary amino acid sequences. Within this platform, we developed a series of local BLAST applications (apps) that take advantage of the speed and sensitivity afforded by BLAST, while allowing for maximal user-definitions and flexible visualization. We tested the efficacy of these apps in control experiments, studying fold-recognition, in which we obtained >90% accuracy in highly divergent sequences (>25% identity). In addition, these same apps were proficient in classifying transmembrane proteins, identifying structural/functional determinants of ion-channels/receptors, and informing structural modeling algorithms. Indeed, these Ada-BLAST informed-structural models were useful in guiding our experimental research on the N-terminus of Transient Receptor Potential ion-channels (TRPs). Taken together, we propose that Ada-BLAST provides a powerful computational tool that is accessible to bench-scientists and computational biologists alike. The codes for Ada-BLAST are publicly available at: http://empathy.rcc.psu.edu/.

Keywords: BLAST, Ada-BLAST, rps-BLAST, PSI-BLAST, twilight-zone, TRP channels, ankyrin repeats, transmembrane prediction, protein function, protein evolution, protein structure, homology modeling, TRPC3, TRPV4, TRP_2, VAMP, SNARE, fusogenic, lipid-binding.

1. Introduction

One of the major challenges that biologists face is identifying the relationships between highly divergent protein sequences. Although many methods (e.g., (12, 16, 17)) have attempted to address the problem, the challenge remains unsolved. In general, when pairwise sequence alignments between protein sequences fall below 25% identity, statistical measurements do not provide support for clear phylogenetic relationships, structural features, or protein function(s) despite intensive research in this area (11, 18-20).

BLAST algorithms are powerful and the institutional "web-based" versions are popular among bench-scientists (21). Although a large amount of highly informative data is collected by BLAST, they are not easily accessed or formatted for use with other algorithms. This is exemplified by performance evaluations for homology detection, where PSI-BLAST underperforms when used as an analysis algorithm (3, 9, 11, 20). Our previous studies suggested that user-defined libraries of Position Specific Scoring Matrices (PSSMs), modifications of BLAST algorithms and settings (e.g. e-value, substitution matrices, sequence-embedding), as well as format changes to BLAST outputs are all useful in increasing the performance of BLAST, particularly for highly-divergent sequences (4, 12, 13, 15, 22).

These findings form the basis of our vision for Adaptive

***Corresponding author:** Randen L. Patterson. Email Adress: rlp25@psu.edu . & Damian B. van Rossum. Email Adress: dbv10@psu.edu .

BLASTing (13), an approach which is designed to maximize user-defined variables and output formats to augment the utility of BLAST data, while maintaining computational ease and efficiency. Ada-BLAST is designed to take any idea from experimentalists and to provide a computational tool to explore his/her idea (Fig. 1). Ada-BLAST begins by identifying protein sequences that contain a relevant structural, functional, and/or evolutionary characteristic to ones question of interest. Following, an rps-BLAST compatible PSSM library of these sequences can be created and amplified using PSI-BLAST. Query sequences of interest can then be aligned with these libraries with user-defined variants of rps-BLAST, each of which has specific advantages (13). We have written simple apps that allow for the various outputs that rps-BLAST provides (e.g. % identity, % coverage, pairwise alignments) to be formatted for use by similarity and distance matrix algorithms, as well as a plethora of visualization and alignment algorithms (6, 13, 14). These format adjustments make Ada-BLAST data highly interoperable, providing a straightforward mechanism for networking multiple algorithms under the umbrella of BLAST. In the present manuscript, we demonstrate the utility that is imparted to BLAST by user-defined applications and suggest that a wealth of information exists in protein sequences that have yet to be fully realized.

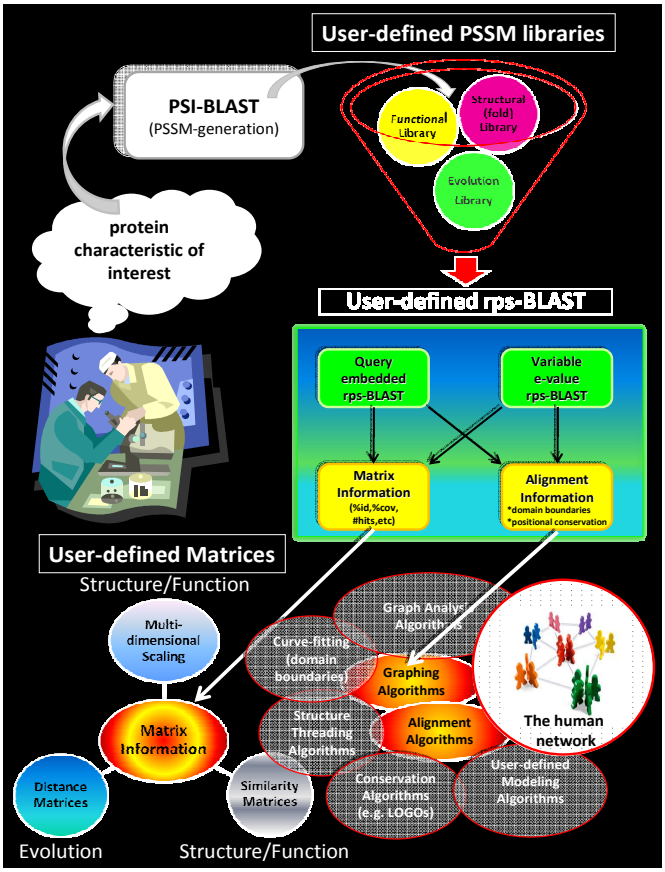


Figure 1. *Ada-BLAST Concept*. This model depicts the logic behind our user-defined interface. The process begins by a scientist deciding on specific protein characteristics that they wish to measure and collecting sequences which are known to have these characteristics. Following, these sequences can be amplified and converted into an rps-BLAST compatible database using PSI-BLAST. Each library can then be entered into a user-defined rps-BLAST interface that allows the user to easily control statistical thresholds and settings. The output from rps-BLAST can then be formatted into either alignment outputs or matrix outputs such that these data are interoperable with a multitude of other analysis and visualization programs.

2. Methods

2.1 *Ada-BLAST Logic*

In our previous studies, we determined that low-identity alignments can be informative to laboratory experiments at multiple scales (e.g., whole protein, single protein domain and single amino acid). We have used these analyses: (i) to reconstruct evolutionary histories (4, 12, 14), (ii) to identify functions in domains of unknown function (4, 23-25), (iii) to classify structural homologues of high sequence divergence (4, 13), and (iv) to inform our biochemical experimentation by isolating key amino acids important to protein function (22-26). We present here the logic behind the user-defined processes in Ada-BLAST platform.

2.2 *User-defined PSSM Libraries*

It is well-established that PSSMs contain more information that single sequences. This is due to the fact that PSSMs contain a frequency distribution of substitutions inherent to homologous yet divergent sequences (3, 5, 20). When large PSSM libraries are constructed, they can be quite powerful; however, it is often the case that institutional PSSM libraries, such as those found on NCBI, are not easily adapted to search for specific protein characteristic. We have created an app that allows for the easy creation of user-defined PSSM libraries that are formatted for use in rps-BLAST. The basic idea underlying our method begins by compiling a set of PSSMs that the query sequence is compared to. These profiles can be obtained from any protein-sequence knowledge-base source (e.g., Protein Data Bank, Pfam, SMART, NCBI Conserved Domain Database (CDD)) (7, 27-29), or they can be locally generated by using PSI-BLAST(5). Our studies demonstrate that when PSSMs are generated using PSI-BLAST (e-value= 10^{-6}, 6 iterations), nearly all of the sequences returned are homologous to the query utilized (6, 12). In this way, a user-defined PSSM library can be rapidly constructed that is enriched for the protein characteristic of interest.

2.3*User-defined rps-BLAST*

Our previous results demonstrate that low-identity alignments obtained from rps-BLAST at insignificant e-values can be informative (as high as e-value=10^{10}) (4, 6, 15). Our results also demonstrate that the removal of e-value using sequence embedding can also be a beneficial local BLAST app (4, 13). Therefore, we have streamlined Ada-BLAST to allow for multiple variations of rps-BLAST to be used, such that multiple datasets can be obtained from the same set of query sequences.

In addition to native rps-BLAST, we utilize embedded alignment algorithm (i.e., embedded rps-BLAST) in our platform. With respect to embedded alignments, we utilize a single domain PSSM database for pairwise comparisons. In brief, we modify the query sequence with a "seed" from the PSSM, creating a consistent initiation site. The "seeds" are generated from the profiles by taking a portion (e.g., 10% in this study, based on the results from our previous studies (12, 13)) of the PSSM sequence (e.g., from the N-terminus or C-terminus). This strategy was designed to amplify and encode the alignments possible for any given query sequence. Instead of a sliding window, we utilized a sliding "seed", a procedure that is similar yet inverse to the embedding strategies employed by Henikoff and Henikoff (1). Since BLAST algorithms are based on a "hit and the extension of the hit" approach, the embedded "seed" creates a consistent initiation site that allows rps-BLAST to extend an alignment even between highly divergent sequences. Next, we filter out the alignments from rps-BLAST using the thresholds of percentage identity and percentage coverage (i.e., the alignment length as a function of the profile length) to remove noisy or random alignments.

Despite the potential of the original embedded alignment strategy for answering a diverse set of biological questions (4, 6, 12, 15), their computational costs made them prohibitively expensive. To address this challenge, we created a novel sequence alignment app that is as sensitive as our previous methods but orders of magnitude faster (see (13) for complete method). Our Ada-BLAST sequence embedding app exploits the similarity among embedded sequences to adaptively avoid expensive computations. Instead of inserting a seed into every position of a query sequence, Ada-BLAST embeds a seed at the query positions where the seed is likely to be extended to an alignment

2.4 *User-defined rps-BLAST Outputs*

Similar to phylogenetic profiles or kernel approaches (30-32), Ada-BLAST can represent a protein as a vector where each entry quantifies the existence of alignments with a given PSSM as measured via alignment algorithm of choice (4, 13). This procedure can be readily adapted to make an unbiased comparison between a series of query sequences by subjecting them to the same screening analysis with the same set of PSSM sequences. Our previous studies determined that % identity, % coverage, and in the case of embedded rps-BLAST, # of hits, are all informative measures which we parse from alignments. In addition, we collect the start and stop position for every alignment above a given threshold which are informative for the detection of domain boundaries (4, 23, 33). We also collect the alignment data for each library to obtain positional data which reflects the conservation of amino-acids from alignments obtained with a given library (see Methods for complete description).

We have developed apps which collect the aforementioned data and provide it in two formats which are highly interoperable with matrix and alignment algorithms. For matrix algorithms, we represent each query sequence (N) as a vector of non-negative numbers (in our case comprised of identity, coverage, and/or # of hits) in M dimensions (M= # of "PSSMs" tested). This N×M data matrix can then be used to create a tree of relationships using hierarchical clustering on the basis of any similarity metrics (e.g. Pearson's correlation (34), multi-dimensional scaling (35), etc) between each query sequence. Importantly, any output obtained from rps-BLAST (e.g. bit score, positional frequency, etc) can easily be incorporated into this matrix using our user-defined apps; however, we have yet to determine the information content of these other variables. In addition, these same data can also be converted to a distance matrix (N×N) using Euclidian distances, making the output appropriate for measuring evolutionary relationships (4, 12, 14).

The alignment information we obtain with our apps allow for alignment boundaries to easily mapped and analyzed using a variety of graphing algorithms. In addition, we have developed an app which translates our data into a WebLogo compatible output (36) for visualization of amino-acid composition (see Methods). Importantly, the outputs presented here are only the ones we have considered and applied; for the computer savvy user, apps for multiple other outputs can be envisioned, coded and easily added to the Ada-BLAST platform.

3. Results

3.1 *Applications for Fold Classification*

As a control for our platform, we performed a fold-recognition assay using 534 sequences from 61 unique fold groups obtained from the "twilight zone" SABmark reference dataset (10). In this dataset, all structurally resolved sequences are <25% identity to one another, making it a challenging dataset. Indeed, our first attempts to resolve this dataset using the NCBI CDD PSSM library obtained an accuracy of only ~30% at a 1 in 100 false-positive rate (4). To test whether user-defined libraries could improve our performance, we generated 61 fold-specific PSSM libraries using 3,995 PDB reference sequences, all of which comprise TZ-SABmark queries in our dataset. These queries contribute, on average, ~3.6 PSSMs for our fold-specific libraries (Fig. S1a). 60 TZ-SABmark sequences that did not create PSSMs at the settings used (see Methods) serve as blind-test sequences (i.e. these sequences do not have self-generated PSSMs in our fold-specific libraries), while other 474 sequences serve as control sequences. All TZ-SABmark sequences were represented as a vector of fold-specific scores (see Methods) to be related in our platform.

In Figure 2a, we present a receiver operating characteristic (ROC)-curve for TZ-SABmark using these libraries at two different thresholds (see Fig. S1b,c for different thresholds). At a false positive rate 0.001, we achieved sensitivity ~0.97 and ~0.94 for e-values of 0.01 and 10^{10} respectively. As expected, the pairwise % identity between TZ-SABmark test sequences and their self-generated PSSMs are low identity (~60% of the alignments are <25% identity). Nevertheless, the

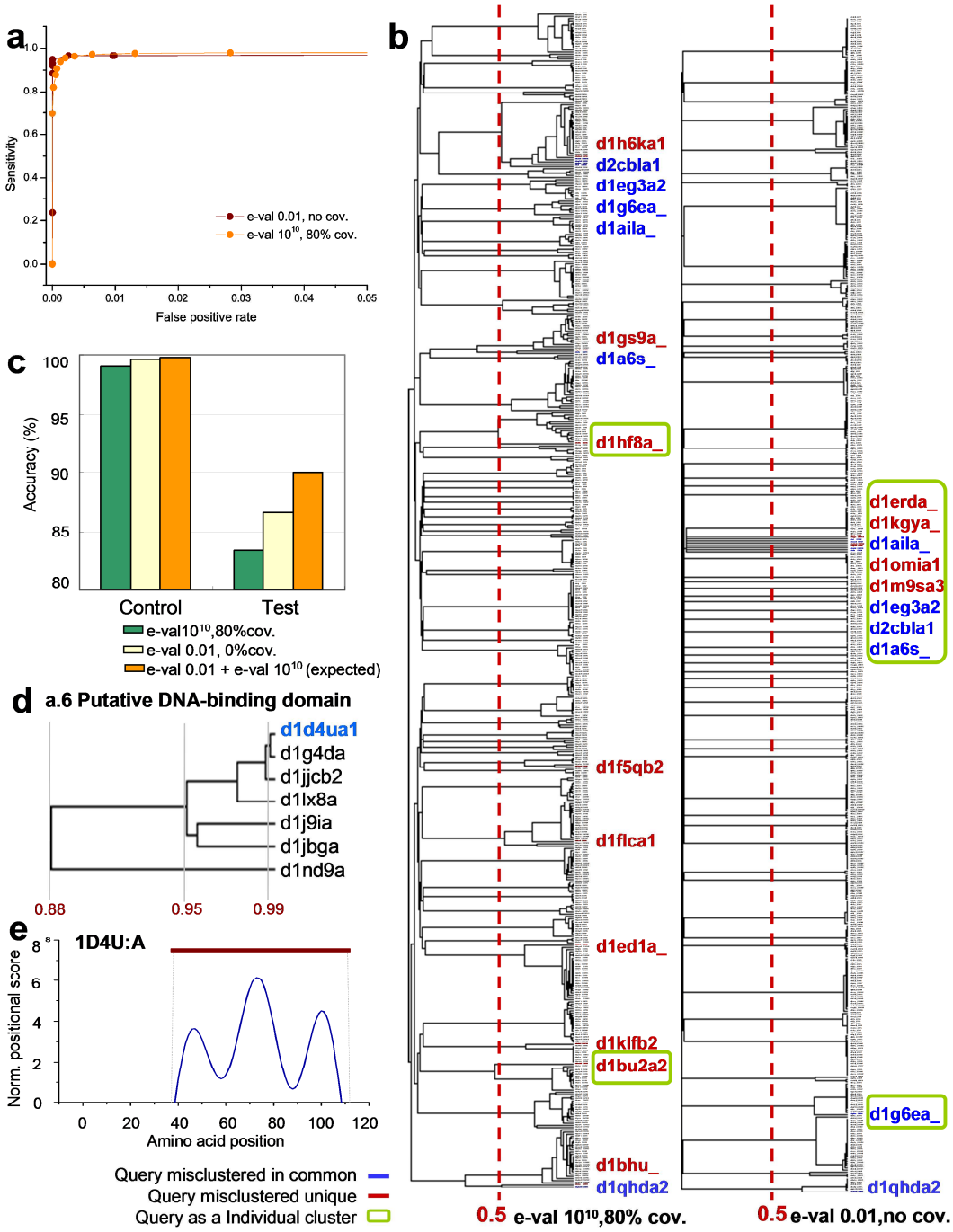


Figure 2. *Fold Recognition and Fold Clustering with the Fold-specific Libraries.* (a) ROC curves of Ada-BLAST. TZ-SABmark queries were encoded in a vector using rps-BLAST alignments of e-value 0.01 and 1010 with 61 fold-specific PSSM libraries (see the Supplemental Methods). (b) Hierarchical clustering of TZ-SABmark queries (left dendrogram: e-value 1010, 80% coverage, right dendrogram: e-value 0.01, no coverage threshold). The queries which could not be clustered with their related folds using Pearson's correlation 0.5 as a cutoff value (red dotted line) are in red (mis-clustered queries exclusively in either dendrogram) or blue (mis-clustered in common). The queries, which could not cluster with any other sequence as forming an individual cluster, are marked in green boxes. (c) Comparison of accuracy of e-value 0.01 and e-value 1010, and expected accuracy when assigning queries into the cluster with higher correlation from e-value 0.01 or 1010 (e-val 0.01 + e-val 1010 expected) for either of 474 test or 60 control TZ-SABmark queries. (d) By hierarchical clustering of TZ-SABmark queries (e-value 1010, 80% coverage) encoded with the 61 fold-specific libraries, d1d4ua1 is correctly clustered with its true fold group without self-generated PSSMs. (e) Predicted a.6 SCOP fold region (blue) in the full length sequence 1D4U:A. Red line annotates an actual Putative DNA-binding domain (a.6) SCOP fold region in the protein. SCOP defined two domains in 1D4U:A, such as d1d4ua1 (a.a.37-111) which is one of TZ-SABmark queries, and d1d4ua2 (a.a.1-36). By SCOP classification, d1d4ua1 is classified as a.6 fold while d1d4ua2 is classified as Glucocorticoid receptor-like (g.39) fold. For regional prediction, Ada-BLAST embedded alignment was run with 10% seed size and 60% coverage and 10% identity thresholds using Putative DNA-binding domain (a.6) fold-specific PSSM library.

pairwise alignments collected with both of the e-value thresholds (0.01 and 10^{10}) for these PSSM libraries are extremely divergent (Fig. S2).

We next sought to determine if hierarchical clustering (34) of TZ-SABmark encoded by our fold-specific PSSM libraries could rebuild appropriate SCOP fold classification (Fig. 2b). Let accuracy be % of TZ-SABmark queries clustered with the sequences from their respective fold groups. We calculated accuracy separately for the control and test datasets (Fig. 2c). For the control sequences, we observe 98.9% and 99.6% accuracy at e-value 10^{10} and 0.01 respectively. If we assign queries into the cluster with higher correlation from e-value 0.01 or 10^{10} (i.e. comparative measurement), we obtain 99.8% accuracy (e-val 0.01+e-val 10^{10} expected in Fig. 2c). For the 60 test sequences, we observe 83.3%, 86.7%, and 90% accuracy at e-value 10^{10}, 0.01, and the comparative measurement respectively. The examples which are properly clustered by comparative measurement are given in Figure S3.

We analyzed TZ-SABmark queries which cannot be clustered with their related fold sequences with Pearson's correlation 0.5 cutoff at e-value 0.01 and 10^{10} respectively (queries in red or blue in Fig 2b). Surprisingly, 9 out of the 10 queries at e-value 0.01 are not clustered with any other sequences (queries in green boxes in Fig 2b-right). It suggests that 99% of the TZ-SABmark queries clustered with at least a single other sequence can be accurately predicted by the fold of the sequences in the same cluster. By comparative measurement between e-value 0.01 and 10^{10} as previously described, we obtain 98.9% accuracy for the entire TZ-SABmark queries because only 6 queries do not correlate with their related fold group clusters in either condition. Importantly, these experiments demonstrate that information collected from traditionally insignificant e-values can be informative.

3.2 *Applications for Transmembrane Protein Classification*

Based on previous results, we next tested whether our matrix outputs can be used to classify transmembrane proteins. All transmembrane proteins are similar in that they have, for the most part, highly hydrophobic helices that transverse lipid-bilayers, yet are all functionally unique. In general, the similarity measurements between two protein sequences are typically done by directly aligning the two sequences, one against the other. However, using Ada-BLAST, we compute an N×M matrix for all query sequences using the same PSSM library (4, 13).

To test the efficacy of our N×M matrices built using embedded alignments or alignments over a range of rps-BLAST e-value thresholds, we curated a set of transmembrane containing proteins from a range of different protein families (e.g. voltage-gated Ca^{2+}, K^{+}, and Cl^{-} channels, calcium-activated K+ channels, cyclic-nucleotide gated channels, transient receptor potential channels (TRPs), receptor tyrosine kinases, G-protein coupled receptors (GPCRs), transporters and exchangers). Each of these 74 sequences was analyzed using both query-embedded and native rps-BLAST using our integral lipid-binding database (ILB-DB) which contains 38,155 PSSMs we generated using PSI-BLAST. The results from this analysis were encoded into an N×M matrix with composite scores (see Methods), and were then subjected to hierarchical clustering as above. In Figures 3a, we report the condition (e-value=100) that achieves the highest degree of classification (see Figure S4,5 for Ada-BLAST embedded and e-value=0.01 clusters, respectively).

While the classifications are not perfect, all three clusters show robust measures for pairing related sequences. Intriguingly, several clades of potassium and cyclic-nucleotide gated channels are close to TRP channels. All three of these groups have 6 TMs and are thought to be structurally related (37). Importantly, these fold/function-specific subgroups are not classification hierarchies in our fold-specific transmembrane library, but emerge as a function of the N×M matrix analysis.

When using Pearson's correlation coefficient 0.5 as a threshold, Ada-BLAST data at e-value=100 achieves 15 clusters (42 proteins) whose proteins have all the same function while embedded data has 12 such clusters (37 proteins). In data not shown, we observe that both the Pearson's correlation values and the overall topology are compromised when measured at e-value=10^{10}. Taken together, this data demonstrates (i) that the classification app has utility, and (ii) that PSSM libraries generated from proteins having a similar characteristic yet diverse function can be used to cluster protein families. Indeed, these results support our previous studies wherein we used PSSM libraries constructed using diverse lipid-binding and nucleic acid-binding folds to identify biologically relevant domains (6, 15, 24, 38).

3.3 *Applications for Identifying Secondary Structural Elements and Amino-acid Conservation*

As a final control, we conducted analyses on a structurally resolved (X-ray Crystallography) transmembrane protein, Bovine Rhodopsin (PDB: 1F88) in order to determine the information content contained in a pure population of embedded alignments and information obtained over a range of e-values using native rps-BLAST (7). Supplemental Figure 11 depicts the output of rps-BLAST (e-value threshold 0.01) for the domain architecture of 1F88. Notably, rps-BLAST returns alignments for multiple PSSMs defined as Serpentine type 7 TM domains. Our theories on structurally/functionally related PSSM libraries predict that additional information below the accepted statistical thresholds can utilized to define, with higher resolution, domain boundaries and secondary structural elements.

This hypothesis was tested and the performance was evaluated against Ada-BLAST and the Hidden Markov Models (TMHMM) (Fig 4a-c, Fig S7) (39). While neither TMHMM nor Ada-BLAST accurately model the entire crystal structure, we observe several interesting features. For example, several of the membrane-spanning helices are interrupted by loop regions that are not identified by TMHMM. Indeed, the C-terminus of 1F88 contains 3 small helices, the last of which is a bent-helix that is believed to be parallel to the membrane

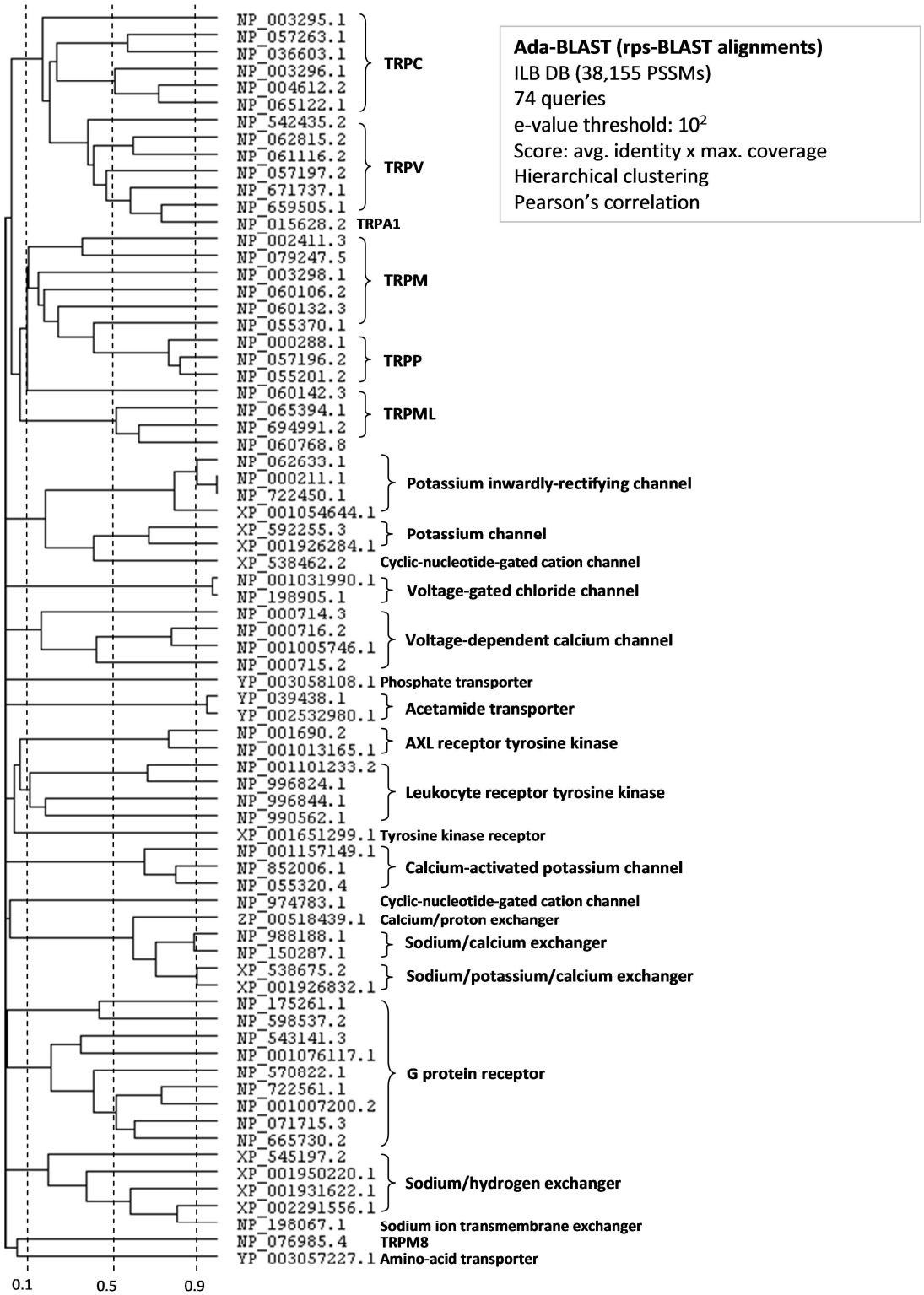


Figure 3. *A Classification based on Heirarchical Clustering.* 74 sequences representing multiple classes of transmembrane containing proteins were hierarchically clustered and visualized by Cluster and Treeview [34]. The dotted lines represent the correlation scores derived from the analysis. Alignments for the ILB-DB PSSMs were derived at e-value=102 threshold (see Supplemental Figure 9-10 for additional clusters of other Ada-BLAST settings).

(aa 288-348, Fig 3c). Various Ada-BLAST settings show multiple views of these smaller helices with embedded data having the highest signal. Another region of interest is contained between aa 91-111, which is a loop in the crystal structure, but is predicted to be a short helix by Ada-BLAST. We theorize that this loop may be, under native conditions, a bent-helix similar to other regions in the protein. Our amino acid conservation data also demonstrates that 6 of the top 8 scor-

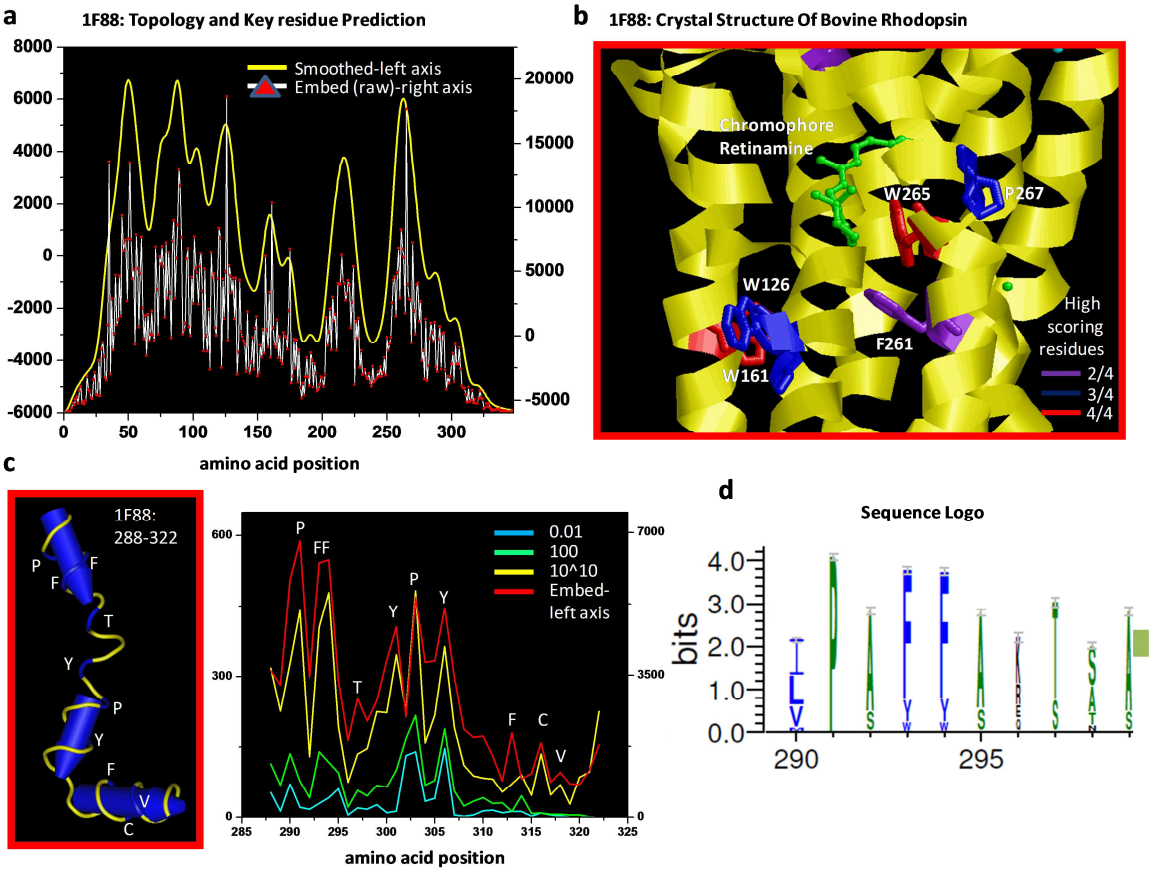


Figure 4. *Ada-BLAST analysis of Bovine Rhodopsin.* (a) The primary amino acid sequence of 1F88 was run against ILB-DB in Ada-BLAST (embedding setting). This graph depicts normalized raw data (right axis) compared with the data smoothed by Fast Fourier Transform (setting=8). These data were normalized by subtrating the average score across the full length sequence from each amino acid position. (see Supplemental Figure 11-12 for rps-BLAST and baseline corrected Ada-BLAST data) (b) Positional Data for TM was compared between Ada-BLAST settings (embedded, e=0.01, e=100, and e=10^{10}). Residues which appeared in 4/4, 3/4, and 2/4 settings are annotated to the structure of 1F88. The majority of the conserved high-scoring residues are in and around the known chromophore binding site. (c) The structure of 1F88 amino acids 288-322 are shown along side the positional data from Ada-BLAST. Interestingly the Ada-BLAST data correlates with transition points in this structural region which bends into three helical subunits. (d) This panel depicts an Ada-BLAST sequence Logo for TM DB (aa 290-299) as an additional view of the postional data.

ing residues (with higher scores representing increased conservation) compared between different Ada-BLAST views surround the chromophore binding site, which must be conserved for vision to occur (Fig 4b). Taken together, our control experiments provide support for the idea that Ada-BLAST can be implemented to investigate any protein sequence, while providing outputs that are interoperable with a variety of analysis algorithms.

3.4 *Using Ada-BLAST to investigate Transient Receptor Potential Channels*

As stated in the introduction, we created the Ada-BLAST interface with bench-scientists in mind. A primary focus of our research group is TRP channel physiology, in particular vertebrate TRPC3, whose physiological role has yet to be clearly elucidated (23, 24, 38, 40). The TRP channel superfamily is well-known for being involved in all areas of sensory perception, as well as variety of other biological functions (41, 42). The TRPC (canonical) channels are most closely related to the TRP channel found in *drosophila* eye which are required for vision (43). Although these channels have been extensively studied, little structural data exists for these channels. This lack of structural data can be confounding when seemingly disparate results are obtained from functional assays where mutational analysis is performed. Further, most of the functional data obtained from cellular studies cannot be predicted using popular institutional sequence analysis algorithms; thus, most TRP channel sequences have little functional annotation as well. When the human TRPC3 sequence is searched against NCBI CDD using rps-BLAST (e-value=0.01), alignments for ankyrin repeats, the ion-channel, and the TRP_2 domain are identified (Fig 5, top). These results are presented in a graphical output that allows for the individual alignments to be observed (Fig 5, middle), and provides links to multiple sources of information for these domains.

While informative, these data cannot easily be extracted in a format that can be used for quantitative comparisons. Conversely, our data outputs provides the positional conservation data obtained using our TM and ankyrin fold-specific data-

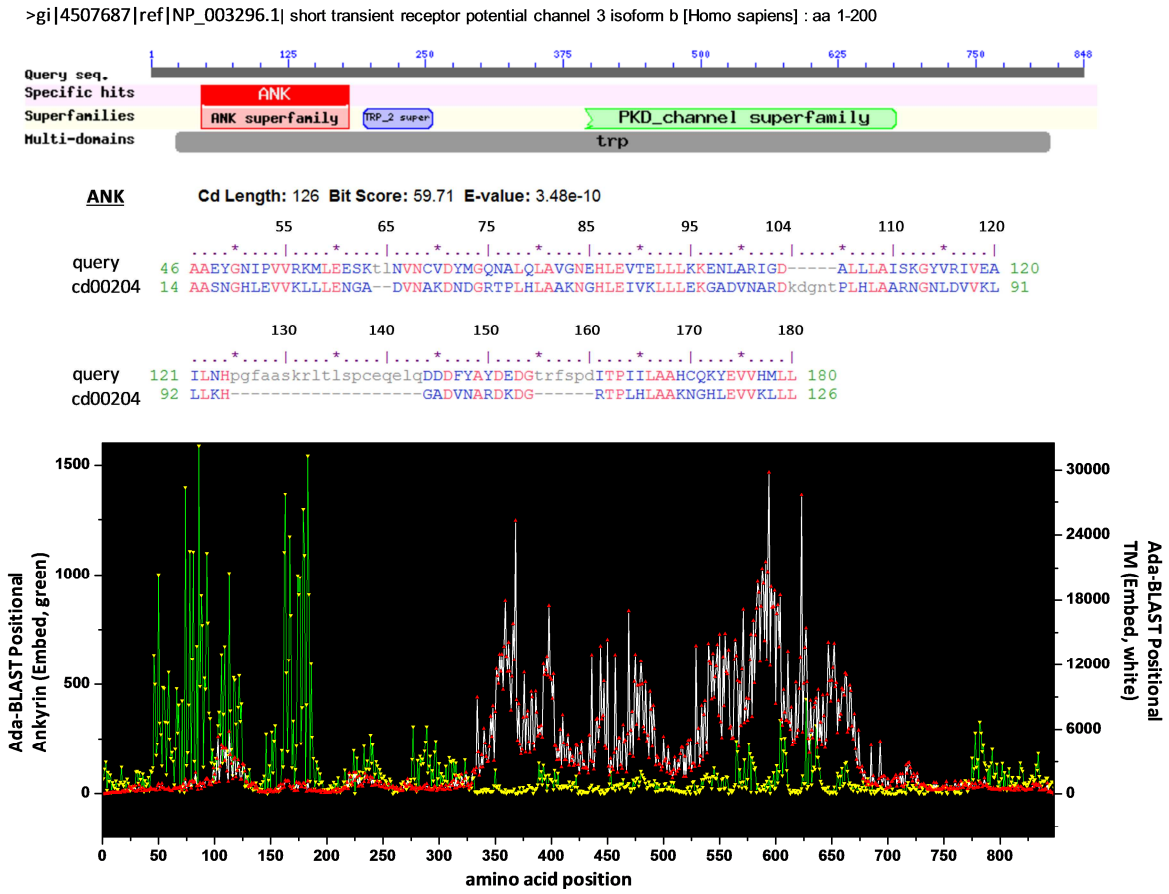


Figure 5. *rps-BLAST and Ada-BLAST analysis of TRPC3.* (*top*) NCBI CDD rps-BLAST output for TRPC3. Three domains are detected as e-value =0.01 (TRP_2, ankyrin, channel). An alignment for each of these domains is also generated. (*bottom)* Ada-BLAST (embedded, rps-BLAST) amino-acid conservation output for TRPC3 measured with ILB DB (white,red) and ANK DB (green,yellow). We observe a clear ankyrin signal in most of the N-terminus of TRPC3, unlike rps-BLAST. In addition, Ada-BLAST predicts that the channel domiain is also underestimated by rps-BLAST, which is supported by TM-HMM (see Supplemental Fig 13).

base (ANK DB, 449 profiles) for TRPC3 (Figure 5, bottom). When compared to NCBI, our embedded app data suggests that the both the channel domain boundaries and the ankyrin repeat boundaries are under-estimated. In Figure S8-9, we compare our ILB-DB results to those obtained using TM-HMM for human TRPC3 and TRPV5. Although the results are very similar, TM-HMM predicts a transmembrane helix in a region of TRPC3 that is known to comprise the TRP-box, which is a peripheral lipid-binding domain, and thus may be an error. In both instances, Ada-BLAST predicts a larger channel region than is annotated by rps-BLAST, which accords with the current ideas in the literature (37).

Similar to our channel boundary predictions, our ANK-DB results suggest that the number of ankyrin repeats in the N-terminus of TRPC3 are underestimated by rps-BLAST (Fig 5, green). To investigate whether this results could be accurate, we compared the results we obtained for TRPC3 with the results we obtained for the structurally resolved ankyrin repeats in TRPV6 (Fig 6a) using our ANK-DB. When the overlapping alignments we obtain are plotted, six clear peaks can be observed, which accord with the 6 ankyrin repeats in the structure (Fig 6b). The amino-acid conservation data (Fig 6a), suggests that repeats 3 and 4 are highly conserved, while repeats 1-2 and 5-6 are less conserved. A very similar pattern for both the overlapping alignment and amino-acid conservation data is obtained for TRPC3 (Fig 6c).

As these alignments extend from amino-acid 1-280 in human TRPC3, this suggests that the TRP_2 domain (~aa 194-260) in TRPC3, which has been implicated in peripheral lipid-binding and vesicle fusion (23, 44), is comprised of ankyrin repeats. By homology, this also suggests that the TRP_2 domain is contained in TRPV channels, which is generally not observed in these channels. To investigate this hypothesis, we aligned the Ada-BLAST-defined region of TRPC3 which was positive for ankyrin alignments with the chicken TRPV4 ankyrin repeat sequence (which has been structurally resolved) using MUSCLE (45). Although the overall alignment is low identity (~13%), the resulting structural model obtained from Modeller (46) using the MUSCLE alignment was high-quality (Fig 6d, 7). The best model we obtained (1470.32458 molpdf score, -21539.84961 DOPE score, and 0.33691 GA341score) was structurally aligned to the TRPV4 structure using MAMMOTH (47) (Fig 6d). The TRPC3 model mirrors the carbon backbone of the TRPV4

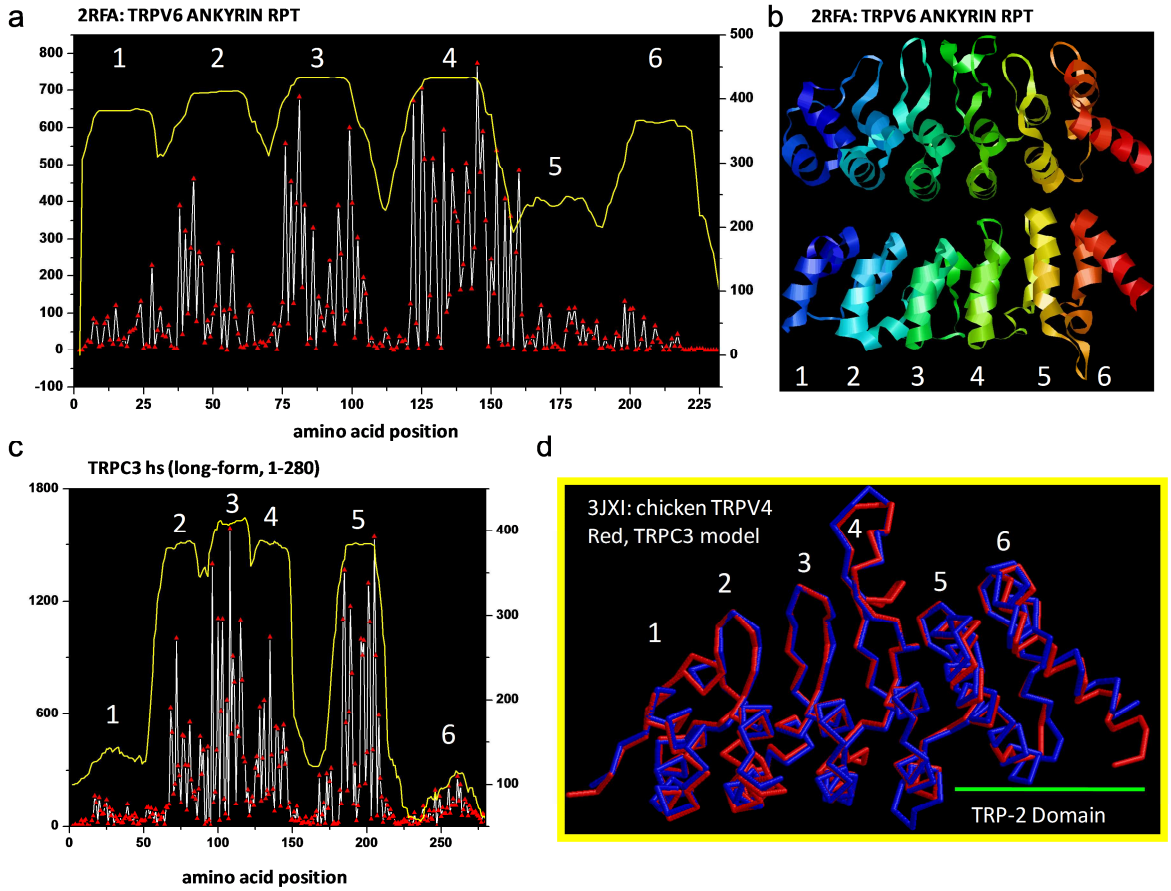


Figure 6. *Anykyrin repeat modeling of TRPC3.* (a) Ada-BLAST ANK ILB measurements (embedded rps-BLAST) of the structurally resolved ankyrin repeats in TRPV6 (*overlapping alignment output*- yellow line, *amino-acid conservation output*- white,red). (b) Two views of the structure of TRPV6 with each ankyrin repeat labeled. (c) Ada-BLAST ANK ILB measurements (embedded rps-BLAST) of TRPC3 (*overlapping alignment output*- yellow line, *amino-acid conservation output*- white,red). (d) Carbon backbone model of TRPC3 (red, Ada-BLAST boundaries) generated with Modeller overlayed with the carbon backbone of TRPV4 (blue). The known TRP_2 domain in TRPC3 comprises the last two ankyrin repeats in this model.

structure in the helical portions of the ankyrin repeats, with the loop regions having a lower correlation; this may be expected if the function of this domain is not conserved between TRPC3 and TRPVs. Of interest, the known TRP_2 domain in TRPC3 corresponds to ankyrin repeats 5 and 6 in the TRPV4 structure (Fig 6d), suggesting that these ankyrin repeats may be involved in the lipid-regulation of TRPVs.

This TRPC3 model allows us to evaluate previous data on this region of the channel with a new perspective. For example, we previously reported that amino-acids 1-48 of TRPC3 comprised a "PH-like" domain, so termed because this region can bind to the partial C-terminal PH-domain (PH-c) contained in phospholipase C-γ (PLCγ) and comprises an intermolecular lipid-binding domain (38). Zhang and colleagues (48) reported that, based on NMR studies, this region does not adopt a PH-fold; a result which our current model supports (Fig. 7). Interestingly, the site we identified as the PLCγ binding site (amino-acids 40-48) comprises an exposed loop/pocket between ankyrin repeats 1 and 2.

Another interesting portion of the TRPC3 N-terminus is contained between amino-acids 121-161 which was demonstrated by us and others to bind SNARE proteins, in particular synaptotagmin (i.e. V-snare, vesicle-associated membrane protein VAMP-1), which regulates channel trafficking (23, 44). These amino-acids in our TRPC3 model create a large loop between ankyrin repeats 3 and 4 (Fig 7). A clear loop resembling a hand can be observed in these loops, which we hypothesize is the binding site for VAMP. When we modeled the T-SNARE binding helix of VAMP into this pocket, we observe that the pocket is sufficiently large to accommodate such a helix; thus it appears this is a reasonable hypothesis to pursue. In addition, the structural model predicts that the VAMP-1 binding pocket is between ~ a.a. 140-152 in TRPC3, which would limit the putative binding region determined by the biochemical data.

4. Experimental Investigation

The aforementioned study by Zhang and colleagues suggested that the region we identified as the binding domain for the PH-c domain of PLCγ was incorrect as they could not achieve binding of this purified fragment with a purified fragment of TRPC3 (a.a. 1-52). Although our study did demonstrate binding to a very similar fragment in yeast-2-hybrid experiments (a.a. 1-48), we used different purified

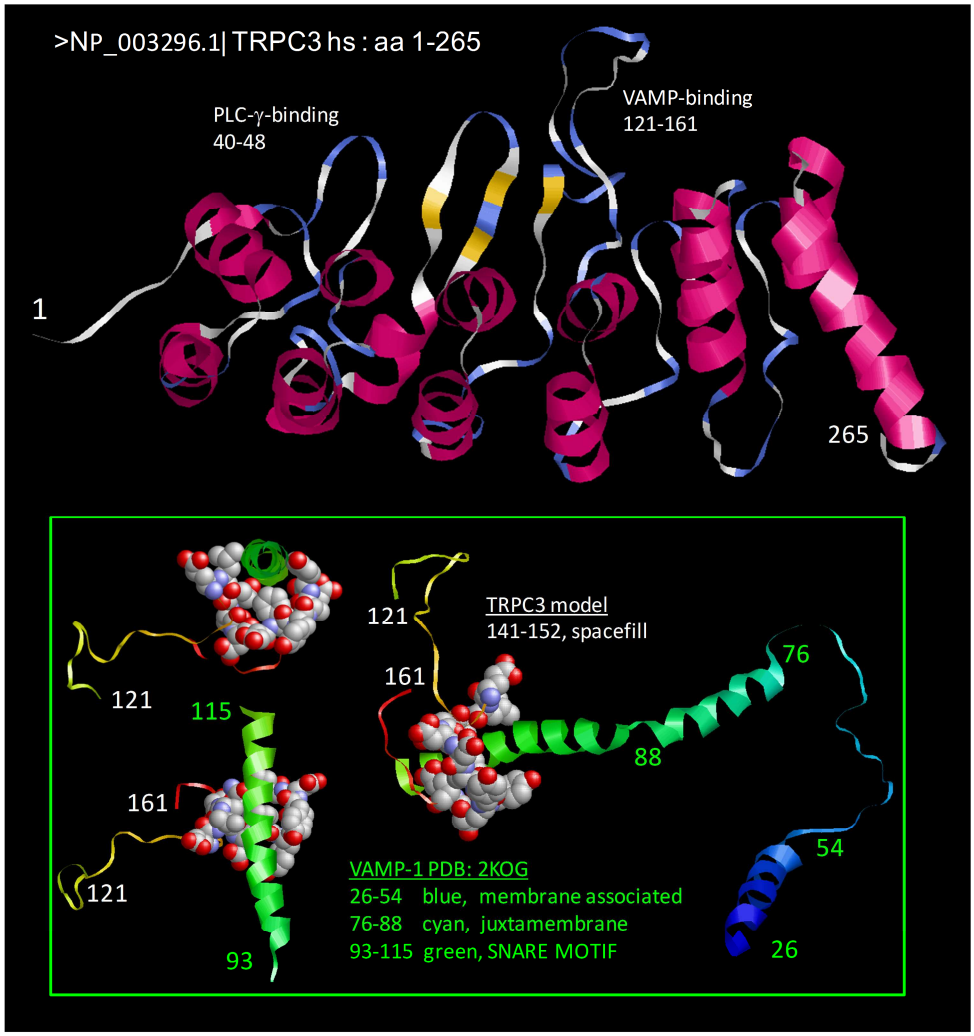


Figure 7. *Homology Model of TRPC3 ankyrin domains.* Amino acids 1-265 of human TRPC3 were aligned to aa 132-384 of chicken TRPV4 (PDB: 3JXI) with Muscle (http://www.ebi.ac.uk/Tools/muscle/). This pairwise alignment was used to make a threaded structural model using MODELLER. This figure shows the TRPC3 homology model colored by structural elements (magenta: Alpha-helices, yellow: Beta-sheets). Key binding pockets for PLC-γ and VAMP-1 are depicted. The model showing the putative interaction between TRPC3 and VAMP-1 (PDB: 2KOG) was generated manually in Deep View PDB viewer based on the known orientation that 2KOG exists in vesicular membranes and the known binding pocket for VAMP-1 in TRPC3 (aa 121-161).

TRPC3 fragments in our in vitro studies (a.a. 40-85 and 1-171). We observe that both of these peptides bind the PH-c domain of PLCγ to form an inter-molecular phosphatidylinositol 4,5-bisphosphate (PIP_2)-binding domain. Further, we demonstrated that a single mutation (F43A), which is at the vertex of a loop in our structural model was capable of alleviating PLCγ binding to these fragments (38) (Fig 8a). Further, we demonstrated that mutation of R14, K16, R18, and R19 also alleviate PIP_2-binding without altering PLCγ PH-c-binding. In our model, these residues form a large solvent accessible positive patch. Based on these observations, we undertook new experimentation to truly determine if this pocket is involved in binding the PH-c of PLCγ to form a PIP_2-binding domain.

We first performed co-immunoprecipitation assays of myc-tagged WT and F43A TRPC3 in HEK-293 cells +/- stimulation with the muscarinic receptor agonist carbachol (CCH, 100μM, 3 minutes). In addition to measuring the binding of TRPC3 with endogenous PLCγ, we also tested binding to the isolated PH-c domain. In the case of the latter, we have over-expressed both WT and F43A TRPC3 with a his-tagged fragment of PLCγ containing the PH-c, as structurally resolved by Zhang and colleagues (a.a. 861-940 of rat PLCγ) (48).

While we observe that while WT TRPC3 readily co-precipitates his-tagged PH-c or endogenous PLCγ we do not observe co-precipitation with the F43A mutant (Fig 8b-c, respectively). We next altered cellular PIP_2 concentrations to determine if PIP_2 is required for this interaction using wortmannin and poly-L-lysine. Low concentrations of wortmannin (1μM) specifically inhibit phosphoinositide-3-kinase (PI3K) and do not deplete PIP_2 levels, while high concentrations (10μM) inhibit a variety of inositol kinases depleting cellular PIP_2 (49). Poly-L-lysine works by chelating PIP_2 such that it is not biologically available (49). We observe that both 10μM wortmannin and poly-L-lysine block WT TRPC3 co-

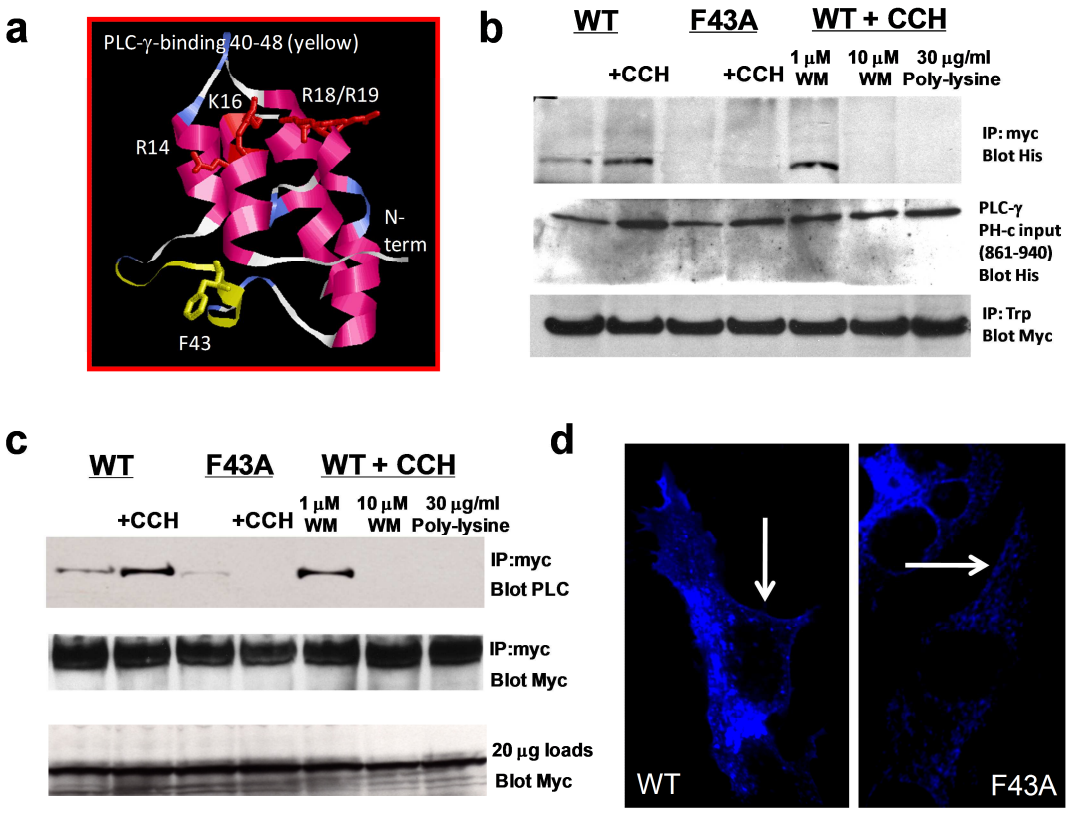


Figure 8. *Experimental validation of the phospholipase C-γpartial PH-domain binding to the ankyrin repeats in TRPC3.* (a) Structural model of the first two ankyrin repeats in TRPC3. Key PLCγ binding residues discovered in (ref) are labeled. (b) Co-immunoprecipitation assays of over-expressed full-length myc-tagged WT and F43A mutants of TRPC3 in HEK-293 cells. (*left*) WT TRPC3 assoaciates with endogenous PLCγ, an interaction that increases upon muscarinic receptor stimulation with carbachol (100μM, 3 minutes). (*right*) WT TRPC3 association with endogenous PLCγ requires PIP_2 as when bioavailable PIP_2 levels are depleted with either 10 μM wortmannin or 30 μg/ml poly-L-lysine, the interaction is lost. (c) Immunostaining of myc-tagged full-length WT and F43A TRPC3 expressed in HEK-293 cells. White arrows demonstrate that WT TRPC3 is enriched near the plasma-membrane, white the F43A mutant is not. (d) Co-immunoprecipitation assays of overexpressed full-length myc-tagged WT and F43A mutants of TRPC3 and his-tagged PLC-γ PH-c (a.a. 861-940 of rat PLCγ) overexpressed in HEK-293 cells. (*left*) WT TRPC3 assoaciates with his-tagged PLC-γ PH-caninteraction that increases upon muscarinic receptor stimulation with carbachol (100 μM, 3 minutes). (*right*) WT TRPC3 association with with his-tagged PLCγ PH-c requires PIP_2 as when bioavailable PIP_2 levels are depleted with either 10μM wortmannin or 30 μg/ml poly-L-lysine, the interaction is lost.

precipitation of overexpressed PH-c or endogenous PLCγ while 1μM wortmannin does not (Fig 8b-c).

We also demonstrated that the F43A mutation abolishes cell-surface expression of TRPC3 (23). To confirm these results, we performed immunostaining in HEK-293 cells for the WT and F43A channels used in this study (Fig. 8d). As expected, we observe that the WT channel is widely distributed throughout the cell, with regions of high concentration that can be observed at the cell periphery (arrow). Conversely, the F43A mutant appears to be completely confined to intracellular vesicles. Taken together, we suggest that a.a. 1-48 of TRPC3 are ankyrin repeats that interact with the PH-c of PLCγ to form an inter-molecular PIP_2-binding domain.

4.1 *TRP_2 domain*

In our studies of the TRP_2 domain in TRPC3 we determined that mutation of S209 and S213 could increase the affinity of this domain for plasma-membrane lipids, although we did not identify residues that could eliminate lipid-binding (23). Through inspecting our model, we identified two arginines (185, 188) in TRPC3 which form a solvent-accessible positively charged patch in the extreme N-terminus of the TRP_2 domain (Figure 9a). We wondered whether these residues were involved in binding to negatively charged plasma-membrane lipids.

To test this hypothesis, we used purified fragments of TRPC3 (amino-acids 161-280) which we have previously demonstrated to bind to liposomes (23). Using PIP-strip© assays as a rapid method for assaying lipid-binding, we observe that WT peptide, and mutant peptides (R185A, R185A/R188A) all bind PIP-strips© and have an identical binding pattern (Fig 9b).

We next tested whether full-length TRPC3 containing the R185A/R188A mutation was expressed in the plasma-membrane using cell-surface biotinylation assays (Fig 9d). We observe that the R185A/R188A mutant is presented on the plasma membrane surface as well, if not better than WT TRPC3. However, when we tested the R185A/R188A for activity by Fura-2 imaging, we observe that these mutations drastically inhibit channel function (Fig 9c). Thus, these mu-

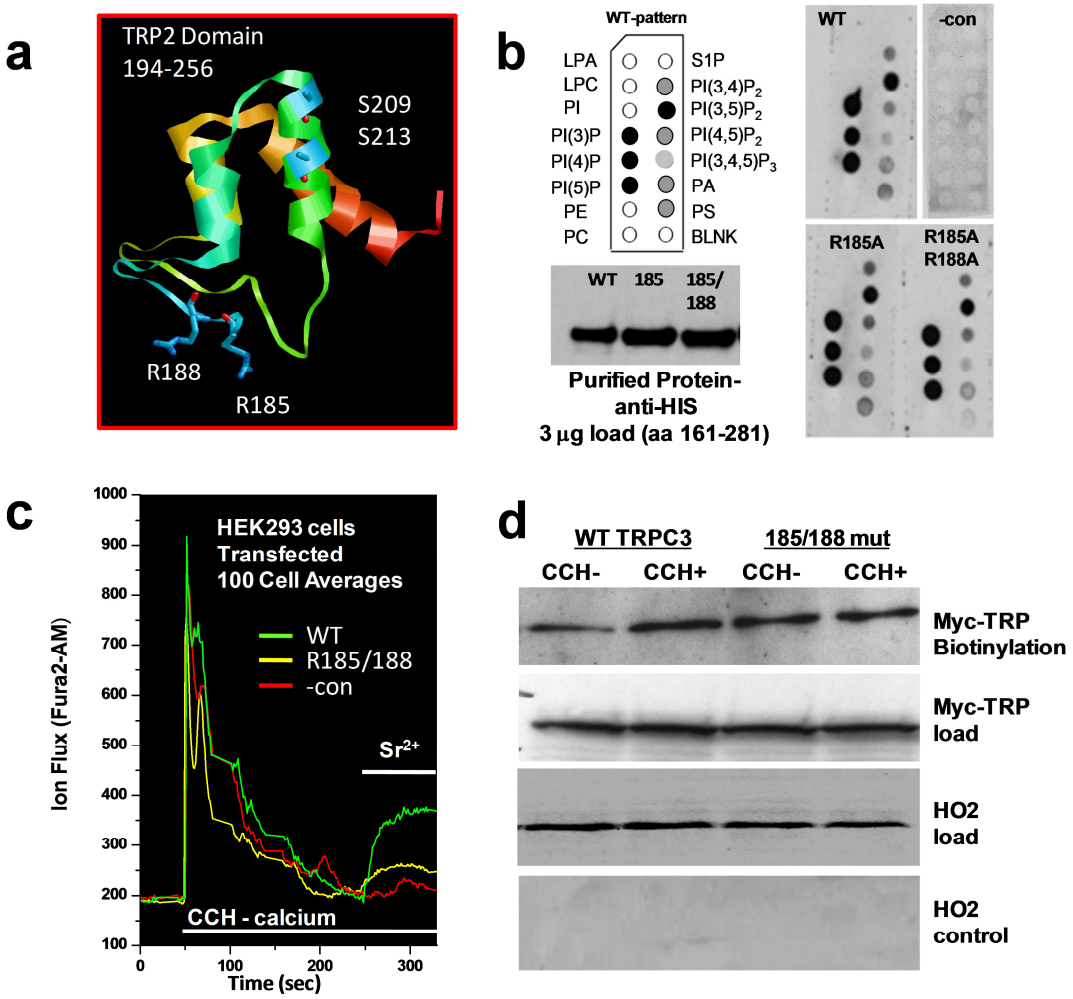


Figure 9. *Experimentally modeling the TRP_2 domain in TRPC3.* (a) Structural model of the first two ankyrin repeats in TRPC3. Key serines (S209, S213) involved in regulating lipid binding discovered in [23] are labeled, as well as two arginines (R185,R188) which we speculated to participate in lipid-binding based on our model. (b) PIP-strip assays (dot Western blots) performed with bacterially purified WT and mutant (R185A, R185A/R188A) TRPC3 (a.a. 161-280) (SDS-PAGE Western blot). We observe no difference in lipid-binding, refuting our speculation. (c) Fura-2AM measurements made in HEK293 cells transfected with YFP alone (-control, red) or transfected with either WT (green) or R185A/R188A (yellow) TRPC3. Cells were acclimated first in nominally Ca^{2+}-free medium, Ca^{2+} pools were released by 100 µM CCH in nominally Ca^{2+}-free medium followed by replacement with CCH and 1 mM Sr^{2+}-containing media. (d) Western blot of biotinylated HEK293 cells (top) and loads (bottom) transfected with either Myc-tagged WT or R185A/R188A TRPC3 alone and stimulated with or without 100 µM carbachol. Input lanes, 20 µg. Anti-HO2 blot serves as an intracellular negative control for biotinylation.

tations do not alter the known functions of the TRP_2 domain, although they are somehow involved in the activation/deactivation of the ion-channel. A second region between a.a. 240-260 of TRPC3 also has a positively charge patch which we are now investigating as a candidate for the lipid-binding activity of the TRP_2 domain (see Supplemental TRPC3 model). In whole, these data demonstrate that the information and models generated using Ada-BLAST provide an excellent resource for performing user-defined analyses toward specific research directions.

5. Discussion

Based on the above results, we propose that the Ada-BLAST algorithm and applications presented in this manuscript provide a highly adaptable user-defined interface for the investigation of primary amino acid sequences. This statement is supported by our observations that: (i) PSSM libraries developed for a specific protein attribute improve the sensitivity and specificity of rps-BLAST, (ii) variations of rps-BLAST that collect low-identity alignments contain information that informs protein structure/function modeling, (iii) modifying the format of rps-BLAST data outputs allows for these data to be interoperable with a plethora of high-performance computational algorithms, and (iv) homology models informed by these data provide a framework to generate hypotheses which can be addressed experimentally. A number of broad implications can be taken from this study.

Systems Biology requires a combination of computational and cellular experiments which obtain quantitative results. The impetus for developing Ada-BLAST was to create an easy-to-use interface that could enhance the information that could be obtained from BLAST algorithms, thereby providing

easily accessible quantitative biological data. As the BLAST algorithm is one of the most popular search algorithms (i.e. the Google of bioinformatics), this is a strong testament to the utility of BLAST. Indeed, BLAST is extremely fast while maintaining a high-level of alignment accuracy (5, 21). By creating apps that allow for a novice computer user to have control over the key components of the rps-BLAST algorithm (PSSM libraries, e-value, output format), this approach allows the bench-scientist to "supercharge" BLAST for investigating protein sequences of interest.

From a biological perspective, having multiple views of a given system is crucial in order to resolve the mechanism of that system. In Ada-BLAST, the data naturally generated by rps-BLAST can be formatted into both alignment and matrix formats; this allows for these data to be visualized in various ways and analyzed by a multitude of other high-performance algorithms, many of which are online resources. Our analysis of the TZ-SABmark dataset demonstrates that when these data are analyzed and visualized using ROC curves, hierarchical clustering, multiple e-values (including embedding), and overlapping alignments, that this multi-view provides confidence for the utility of fold-specific PSSM libraries.

Similarly, when this "multi-view" approach is applied to ion channels, it provides models which are highly testable. The Ada-BLAST model for the N-terminus of TRPC3 suggests that: (i) the number of ankyrin repeats were previously underestimated (Fig 5-6), (ii) a.a.1-260 of human TRPC3 is homologous to a.a.132-384 of chicken TRPV4 (~13% pairwise identity) (Fig 6-7), (iii) that the PH-c of PLCγ likely interacts with TRPC3 through binding a soluble loop between ankyrin repeats 1 and 2, (iv) the membrane-fusogenic TRP_2 domain in TRPC3 is comprised of ankyrin repeats, and (v) that TRPC3 likely binds VAMP via a large loop between ankyrin repeats 3 and 4. Our biochemical experiments confirm that the PH-c of PLCγ forms an inter-molecular lipid-binding domain. Further, this interaction is required for channel function and that R185 and R188, which are proximal to the TRP_2 domain, do not bind lipid, but do regulate channel function.

Our results also support the idea that statistical thresholds are often too stringent in domain detection algorithms. For example, rps-BLAST does not report a channel domain alignment in human TRPV5 channel (gi|22547180) at statistical limits. In this study, we found that additional information contained in alignments well below accepted statistical thresholds can be utilized to identify domain boundaries and secondary structural elements. Future analysis on a sufficiently large data set is required to identify and optimize the multiple variables that can identify highly divergent yet informative alignments. Nevertheless, we propose that there is a wealth of information below statistical values that can aid researchers in annotating protein structure/function.

To these ends we recently completed the CASP9 structural competition (50, 51) with Ada-BLAST using a PSSM library constructed using the most recent SCOP structural database (52). We expect that the results from the CASP experiment will have an immediate impact on our ability to construct PSSM libraries of high-utility. In conclusion, we propose that future work aimed at (i) creating comprehensive and refined PSSM libraries and (ii) exploring sequence embedding at the level of the PSSM (COBBLER (1)) and within the query (query-embedded Ada-BLAST), may have the ability to exponentially increase the functional annotation of all classes of proteins across taxa.

6. Supplementary material

Supporting Information Available: *Supplemental Methods. Supplemental Figure 1.* Fold recognition performance of Ada-BLAST with different settins given fold-specific libraries. *Supplemental Figure 2.* Characterization of alignments used by Ada-BLAST at e-value 0.01 and 10^{10} thresholds. *Supplemental Figure 3.* Comparison of Ada-BLAST dendrograms of e-value 0.01 and 10^{10} thresholds. *Supplemental Figure 4-5.* Hierarchical clustering of transmembrane containing proteins with additional Ada-BLAST settings. *Supplemental Figure 6.* rps-BLAST analysis of 1F88. *Supplemental Figure 7.* The characterization of structural elements in 1F88. *Supplemental Figure 8.* Ada-BLAST comparison with TMHMM and rps-BLAST for TRPC3. *Suppelemental Figure 9.* Ada-BLAST comparison with TMHMM and rps-BLAST for TRPV5. *Supplemental Files.* Homology models of TRPC3 a.a. 1-265 and a.a. 121-161 that are generated using TRPC4 and VAMP-1 as template structures, respectively.

http://www.jiomics.com/index.php/jio/rt/suppFiles/33/0

Acknowledgements

This work was supported by the Searle Young Investigators Award and start-up money from PSU (RLP), NCSA grant TG-MCB070027N (RLP, DVR), The National Science Foundation 428-15 691M (RLP, DVR), and The National Institutes of Health R01 GM087410-01 (RLP, DVR). This project was also funded by a Fellowship from the Eberly College of Sciences and the Huck Institutes of the Life Sciences (DVR) and a grant from the Pennsylvania Department of Health using Tobacco Settlement Funds (DVR). The Department of Health specifically disclaims responsibility for any analyses, interpretations or conclusions. We would especially like to thank Jason Holmes and the CAC center for their superior support. We would like to thank Dr. Jaewoo Kang and Dr. Dongwon Lee for their contribution to develop the heuristic embedded alignment algorithm. We would also like to thank Drs. Robert E. Rothe, Jim White, Peter J. Hudson, Jenji Cohan, Barbara Van Rossum, Max Moon, Anderson Sharer, and J. Hendrix for creative dialogue.

References

1. S. Henikoff, J. G. Henikoff. *Protein Sci.* 6 (1997) 698-705.
2. M. Gribskov, A. D. McLachlan, D. Eisenberg. *Proc. Natl. Acad. Sci. U. S. A* 84 (1987) 4355-4358.

3. L. Jaroszewski, L. Rychlewski, Z. Li, W. Li, A. Godzik. *Nucleic Acids Res.* 33 (2005) W284-W288.
4. K. D. Ko, Y. Hong, G. S. Chang, G. Bhardwaj,. D. van Rossum, R. L. Patterson. *Physics Archives* (2008) arXiv:0806.239, q-bio.Q.
5. S.F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman. *Nucleic Acids Res.* 25 (1997) 3389-3402.
6. K. D. Ko, Y. Hong, G. Bhardwaj, T. M. Killick, D. B. van Rossum, R. L. Patterson. *Physics Archives* (2009) arXiv:0911.0652v1, q-bio.QM 1-21.
7. A. Marchler-Bauer, J. B. Anderson, P. F. Cherukuri, C. Weese-Scott, C., L. Y. Geer, M. Gwadz, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, *et al. Nucleic Acids Res.* 33 Database Issue (2005) D192-D196.
8. C. Z. Cai, L. Y. Han, Z. L. Ji, X. Chen, X, Y. Z. Chen. *Nucleic Acids Res.* 31 (2003) 3692-3697.
9. J. Soding. *Bioinformatics.* 21 (2005) 951-960.
10. W. I. Van, I. Lasters, L. Wyns. *Bioinformatics.* 21 (2005) 1267-1268.
11. G. Yona, M. Levitt. *J. Mol. Biol.* 315 (2002) 1257-1275.
12. G. S. Chang, Y. Hong, K. D. Ko, G. Bhardwaj, E. C. Holmes, R. L. Patterson, R. L., D. B. van Rossum. *Proc. Natl. Acad Sci U. S. A* 105 (2008) 13474-13479.
13. Y. Hong, D. Lee, J. kang, D. B. van Rossum, R. L. Patterson. *Physics Archives* (2009) arXiv:0911.0650v1, q-bio.QM, 1-21.
14. G. Bhardwaj, Z. Zhang, Y. Hong, K. D. Ko, G. S. Chang, E. J. Smith, L. A. Kline, D. N. Hartnraft, E. C. Holmes, R. L. Patterson, *et al. Physics Archives* (2010) q-bio.PE, q-bio.QM, 1-13.
15. Y. Hong, D. Chalkia, K. D. Ko, G. Bhardwaj, G. S. Chang, D. B. van Rossum, R. L. Patterson. *Journal of Proteomics and Bioinformatics* (2009) 139-149.
16. J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, C. Chothia. *J. Mol. Biol.* 284 (1998) 1201-1210.
17. J. D. Blake, F. E. Cohen. *J. Mol. Biol.* 307 (2001) 721-735.
18. L. Wang, T. Jiang. *J. Comput. Biol.* 1 (1994) 337-348.
19. T. Rognes, E. Seeberg. *Bioinformatics* 14 (1998) 839-845.
20. A. A. Schaffer, Y. I. Wolf, C. P. Ponting, E. V. Koonin, L. Aravind, S. F. Altschul. *Bioinformatics* 15 (1999) 1000-1011.
21. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman. *J. Mol. Biol.* 215 (1990) 403-410.
22. A. K. Mustafa, D. B. van Rossum, R. L. Patterson, D. Maag, J. T. Ehmsen, S. K. Gazi, A. Chakraborty, R. K. Barrow, L. M. Amzel, S. H. Snyder. *Proc. Natl. Acad. Sci. U. S. A*. (2009)
23. D. B. van Rossum, D. Oberdick, Y. Rbaibi, G. Bhardwaj, R. K. Barrow, N. Nikolaidis, S. H. Snyder, K. Kiselyov, R. L. Patterson. *J. Biol. Chem.* 283 (2008) 34384-34392.
24. G. Caraveo, D. B. van Rossum, R. L. Patterson, S. H. Snyder, S. Desiderio. *Science* 314 (2006) 122-125.
25. N. C. Zachos, D. B. van Rossum, X. Li, G. Caraveo, R. Sarker, B. Cha, S. Mohan, S. Desiderio, R. L. Patterson, M. Donowitz. *J. Biol. Chem.* (2009)
26. [26] A. Chakraborty, M. A. Koldobskiy, K. M. Sixt, K. R. Juluri, A. K. Mustafa, A. M. Snowman, D. B. van Rossum, R. L. Patterson, S. H. Snyder. *Proc. Natl. Acad Sci U. S. A* 105 (2008) 1134-1139.
27. [27] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi. *J. Mol. Biol.* 112 (1977) 535-542.
28. [28] I. Letunic, R. R. Copley, S. Schmidt, F. D. Ciccarelli, T. Doerks, J. Schultz, C. P. Ponting, P. Bork. *Nucleic Acids Res.* 32 Database issue (2004) D142-D144.
29. [29] E. L. Sonnhammer, S. R. Eddy, R. Durbin. *Proteins* 28 (1997) 405-420.
30. [30] J. P. Vert. *Bioinformatics.* 18 Suppl 1 (2002) S276-S284.
31. [31] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates. *Proc. Natl. Acad Sci U. S. A* 96 (1999) 4285-4288.
32. [32] Y. Kim, S. Subramaniam. *Proteins* 62 (2006) 1115-1124.
33. [33] K. D. Ko, G. Bhardwaj, Y. Hong, G. S. Chang, K. Kiselyov, D. B. van Rossum, R. L. Patterson. *Commun. Integr. Biol.* 2 (2009) 133-137.
34. [34] M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein. *Proc. Natl. Acad. Sci. U. S. A* 95 (1998) 14863-14868.
35. [35] G. Kemmler, B. Holzner, M. Kopp, M.. Dunser, R. Greil, E. Hahn, B. Sperner-Unterweger. *Qual. Life Res.* 11 (2002) 223-233.
36. [36] G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner. *Genome Res.* 14 (2004) 1188-1190.
37. [37] F. Yang, Y. Cui, K. Wang, J. Zheng. *Proc. Natl. Acad. Sci. U. S. A* 107 (2010) 7083-7088.
38. [38] D. B. van Rossum, R. L. Patterson, S. Sharma, R. K. Barrow, M. Kornberg, D. L. Gill, S. H. Snyder. *Nature* 434 (2005) 99-104.
39. E. L. Sonnhammer, H. G. von, A. A. Krogh. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6 (1998) 175-182.
40. R. L. Patterson, D. B. van Rossum, D. L. Ford, K. J. Hurt, S. S. Bae, P. G. Suh, T. Kurosaki, S. H. Snyder, D. L. Gill. *Cell* 111 (2002) 529-541.
41. K. Venkatachalam, C. Montell. *Annu. Rev. Biochem.* 76 (2007) 387-417.
42. D. E. Clapham. *Nature* 426 (2003) 517-524.
43. C. Montell, G. M. Rubin. *Neuron* 2 (1989) 1313-1323.
44. B. B. Singh, T. P. Lockwich, B. C. Bandyopadhyay, X. Liu, S. Bollimuntha, S. C. Brazer, C. Combs, S. Das, A. G. Leenders, Z. H. Sheng *et al. Mol Cell* 15 (2004) 635-646.
45. R. C. Edgar. *Nucleic Acids Res.* 32 (2004) 1792-1797.
46. N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, A. Sali. *Curr. Protoc. Protein Sci* Chapter 2, Unit, 2007.
47. A. R. Ortiz, C. E. Strauss, O. Olmea. *Protein Sci.* 11 (2002) 2606-2621.
48. W. Wen, J. Yan, M. Zhang. *J. Biol. Chem.* 281 (2006) 12060-12068.
49. D. Kong, T. Yamori. *Curr. Med. Chem.* 16 (2009) 2839-2854.
50. K. Karplus, S. Katzman, G. Shackleford, M. Koeva, J. Draper, B. Barnes, M. Soriano, R. Hughey. *Proteins* 61 Suppl 7 (2005) 135-142.
51. K. T. Simons, R. Bonneau, I. Ruczinski, D. Baker. *Proteins* Suppl 3 (1999) 171-176.
52. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia. *J. Mol. Biol.* 247 (1995) 536-540.
53. A. Bradley*Pattern Recognition* 30 (1995) 1145-1159.
54. A. Sali, L. Potterton, F. Yuan, V. H. van, M. Karplus. *Proteins* 23 (1995) 318-326.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.25

Proteomic analysis of Chinese kale (*B. alboglabra*) leaves during growth

Zuraifah Asrah Mohamad[1], Daranee Chokchaichamnankit[2], Kisana Bhinija[3], N. Monique Paricharttanakul[2], Jisnuson Svasti[2,4], Pattana Srifah Huehne[3], Chantragan Srisomsap*[2].

[1]Applied Biological Sciences Program, Chulabhorn Graduate Institute, Bangkok 10210, Thailand; [2]Laboratory of Biochemistry, Chulabhorn Research Institute, Bangkok 10210, Thailand; [3]Laboratory of Biotechnology, Chulabhorn Research Institute, Bangkok 10210, Thailand; [4]Department of Biochemistry, Faculty of Science, Mahidol University, Bangkok 10400, Thailand.

Received: 9 July 2010 **Accepted:** 23 September 2010 **Available Online**: 1 October 2010

ABSTRACT

Brassica alboglabra (Chinese kale) is a vegetable extensively grown in Thailand, which has high nutritional value and useful phytochemicals. Farmers generally harvest *B. alboglabra* starting from the fifth week of growth to sell in the market. In this study, changes in protein expression during growth and development of *B. alboglabra* were investigated. Proteins were extracted from two to eight-week leaves, and a total of 334 protein spots separated by two-dimensional gel electrophoresis and selected 103 spots were digested and analyzed by using LC-MS/MS. The identified proteins could be classified into nine classes, namely proteins involved in photosynthesis and photorespiration, amino acid metabolism, carbon-compound and carbohydrate metabolism, protein metabolism, stress response, cellular communication and signal transduction, glycolysis and gluconeogenesis, unidentified and others. The highest number of proteins was the proteins involved in photosynthesis and photorespiration, presumably because leaves are the primary sites for photosynthesis and photorespiration, so there is an induction of proteins such as ribulose bisphosphate carboxylase and ribulose bisphosphate carboxylase activase. This is the first study to investigate protein expression in *B. alboglabra* leaves during growth and development. The studies provide information for protein database in this plant species.

Keywords: Proteomic; Chinese kale; *Brassica alboglabra*; Leaves; Growth.

1. Introduction

Plants are unique in possessing the ability to generate own energy from photosynthesis and carbon fixation, which eventually provide food and oxygen to the other organisms including humans (Micol, 2009). Leaves contain chlorophyll, a molecule which acts as a photoreceptor to trap sunlight for the production of oxygen and simple sugars, and has been suggested to have anti-mutagenic and anti-carcinogenic properties, as well as protective properties against chronic diseases such as coronary heart diseases, diabetes, and cataracts (Ma and Dolphin, 1999). Besides, photosynthesis provides the carbon skeleton for the plant metabolic pathways to support the plant growth and development.

The "Omics" era enables researchers to study plant development in-depth by using novel tools such as transcriptomics, proteomics, metabolomics and modelling (Hennig, 2007). Proteomic studies allow the detection and identification of proteins expressed, leading to the discovery of genes and pathways involved in stress responsiveness and tolerance (Salekdeh *et al*, 2007).

Several studies on proteomic analyses related to plant growth and development processes have been reported, including rice, soybean, wheat, barley, maize, and Arabidopsis (Nozu *et al*. 2006, Mooney *et al*. 2006, Salekdeh *et al*. 2007). Proteomics is a sensitive tool for molecular studies geared towards the identification and characterization of genes responsible for regulating key plant processes and traits. Proteomic studies of plants have been extensive, since entire genome of the reference plant species *Arabidopsis thaliana* has been sequenced. *Arabidopsis thaliana*, in the family of Brassicaceae genus, has been used as a model organism in molecular biology. Plant proteomics studied in several food crops such as rice (Goff, 2002) and wheat (Ven-

***Corresponding author:** Chantragan Srisomsap. Laboratory of Biochemistry, Chulabhorn Research Institute, 54 Moo 4 Vibhavadee Rangsit Road, Laksi, Bangkok 10210, Thailand. Phone: 66-2-5740622 ext. 3715. Fax: 66-2-5740622 ext. 3716. Email Address: chantragan@cri.or.th

sel *et al*. 2005), vegetables such as tomatoes (Chen, 1999) and peas (Kav, 2005), and medicinal plants such as ginseng (Lum *et al*. 2002). Proteomics also serves as a modern approach to ascertain the protein alterations due to stress or the response of tolerance to stress environment (Salekdeh *et al*. 2007).

Chinese kale or Chinese broccoli (*Brassica alboglabra*) is one of the most commonly grown leafy vegetables in the Southeast Asian region and Thailand, as it tolerates the tropical weather, in particular which makes it suitable to be grown throughout the year (Issarakraisila *et al*. 2007). Chinese kale is categorized under the *Brassica oleracea* group, the same species with broccoli and curly kale. Its edible parts consist of waxy, broad and dark green leaves and thick and crunchy stems. The slightly bitter and crunchy taste, with high nutritional value includes protein, calcium, iron, vitamins, and fiber, have made Chinese kale a favorite among the vegetables in the region. Studies have revealed that Brassica vegetables contain phytochemicals such as glucosinolates to fight against cancer (van Poppel *et al*. 1999, Talalay *et al*. 2001, La *et al*. 2009). Chinese kale is harvested when it is still immature starting from the fifth week and can be stored for about 7-10 days at 0ºC, with relative humidity 95-100% (Sukontasing, 2009).

Despite its' popularity and nutritional value, proteomics of Chinese kale has not yet been established. Therefore this study of the protein alterations in *Brassica alboglabra* leaves at two to eight weeks of growth provides a proteomic database on this plant species and allows comparison of the protein changes over time.

2. Material and methods

2.1 Plant materials

The seeds for the *B. alboglabra* were obtained from a local seed company. The seeds for the *B. alboglabra* were grown in a pot of 8 cm in diameter filled with decomposed soil under the open environment at Chulabhorn Research Institute compound. The plants were supplied with organic and chemical fertilizers only on day 0. The plants were harvested in the morning prior to extraction, and the leaves were cut and weighed before subjecting to protein extraction.

2.2 Time course study

The plants were grown with no other treatment except water and harvested on the second, third, fourth, fifth, sixth, seventh, and eighth weeks of growth. Photographs were taken at every point of time. Plants were harvested for leaves for protein analysis and identification. The time course study was performed twice. The protein analysis from leaves were subjected onto two-dimensional electrophoresis 3-5 sets.

2.3 TCA/ acetone precipitation

0.1 g plant samples were ground in liquid nitrogen and resuspended in 1 mL solution of 10% TCA and 0.07% DTT in cold acetone. The samples were incubated at 20^0C for 1 hour, followed by centrifugation at 14 000 rpm for 10 minutes. The supernatant was discarded, and the pellet was resuspended in 1 mL of 0.07% DTT in cold acetone, incubated and centrifuged as previously described. This step was repeated twice. Then, the pellet was vacuum-dried for 30 minutes and resuspended in 500 µL of rehydration buffer consisted of 8 M urea, 2% CHAPS, 40 mM DTT, 0.2% ampholytes 3-10, and 2 mM tributylphosphine (TBP). The samples were then vigorously vortexed for 5 minutes and incubated at 4 ^{0}C overnight. The next day, the samples were centrifuged as previously described. The supernatant was collected for protein determination by Bradford assay.

2.4 Two-dimensional electrophoresis

The two-dimensional electrophoresis was performed using Immobiline™ DryStrip pH 3-10 non-linear, 7 cm (GE Healthcare). The 125 µL solution containing 150 µg of protein and rehydration buffer was loaded and kept at room temperature overnight. The first IEF was conducted at 7,000 Vhrs, 55 µA/strip and run for 6 hours. Prior to the second dimension SDS-PAGE gels, the strips were soaked with equilibration buffer 1 (0.5 M Tris-HCl pH 6.8, 6 M urea, 30% glycerol, 1% SDS, and 1% DTT) with gentle agitation at room temperature for 10 minutes followed by soaking in equilibration buffer 2 (0.5 M Tris-HCl pH 6.8, 6 M urea, 30% glycerol, 1% SDS, and 2.5% iodoacetamide). The IPG strips were embedded within molten agarose directly on top of a 1 mm 12.5 % SDS-PAGE gel. The SDS-PAGE was performed at 12 mA for 3 hours.

2.5 Protein staining with Coomassie blue R-250

The gels were stained with 0.1% Coomassie blue R-250, 40% methanol and 10% acetic acid overnight, and then destained with 40% methanol and 5% acetic acid for 3 hours, followed by 10% of methanol and 5% acetic acid for 2 hours.

2.6 Gel Scanning and Image analyses

The gels were scanned by an ImageScanner II (GE Healthcare, Uppsala, Sweden). The images were analysed with the Image Master 2D Platinum 6.0 (GE Healthcare) by matching and comparing the different percent volume (% Volume) of protein spots. A master gel was selected from the gel with the highest number of spots. Significant change refers to ≥2 fold change of percent volume. Descriptive data were analyzed using Graphpad Prism 5 Software to obtain Means ± SD.

2.7 Protein identification by LC/MS/MS

The spots from the 2-DE gels were excised with a clean scalpel into 1 mm X 1 mm cubes and transferred to a 0.5 ml microfuge tube followed by washing 2 times with RO water. The samples underwent the steps of destaining with 0.1M NH_4HCO_3/50% acetonitrile, reduction of disulphide bonds with 0.1M NH_4HCO_3/10mM DTT/1mM EDTA, alkylation with 100mM iodoacetamide/0.1M NH_4HCO_3 and digestion with trypsin. The LC/MS/MS analyses were carried out using

a capillary LC system (Waters) coupled to a Q-TOF mass spectrometer (Micromass, Manchester, UK) equipped with a Z-spray ion-source working in the nanoelectrospray mode. Glu-fibrinopeptide was used to calibrate the instrument in MS/MS mode. The tryptic peptides were concentrated and desalted on a 75 µm ID X 150 mm C18 PepMap column (LC Packings, Amsterdam, Netherlands). Eluents A and B were 0.1 % formic acid in 97% water, 3% acetonitrile and 0.1% formic acid in 97% acetonitrile respectively. Six µl sample was injected into the nanoLC system, and separation was performed using the following gradient: 0 min 7% B, 35 min 50% B, 45 min 80% B, 49 min 80% B, 50 min 7% B, 60 min 7% B. The database search from SWISS-PROT and NCBI was performed with ProteinLynx,. The Mascot search tool available on the Matrix Science site screening (http://www.matrixscience.com) was used for some proteins which were not available in previous database. Protein function was obtained from UniProtKB website.

3. Results and Discussion

In this study, *B. alboglabra* seeds were grown and plants were harvested from two weeks until eight weeks. The growth and development were observed from the first week (seedling) to eighth week. There were changes in the growth observed from the seedling to the eighth week as shown in Figure 1. At the seedling stage, plants possessed only cotyledons and true leaves were observed starting from the second week of harvest. The leaves began maturing at the third week of growth with waxy, dark green surface appearances observed, while the stems gradually became hard throughout the time course. The cotyledons began to wilt at the seventh week of growth.

Plants from two to eight weeks were harvested in the morning, leaves were cut and proteins were extracted for 2-DE analyses. In the first dimension, proteins (150 µg) were separated by using IEF strips of pH 3-10, followed by the second dimension of 12.5 % SDS PAGE and stained with Coomassie blue. The gels of the second to the eighth week leaves were scanned and shown in Figure 2. From the second to the eighth week of 2-DE protein patterns, the lowest number of proteins was observed in the second week and began to increase towards the sixth week and decreased at the seventh and eighth weeks. This may indicate increase of protein expression during the growth and development of the *B. alboglabra* from the second to the sixth week and decline starting the seventh week. There may also be morphological changes of the leaves which became thicker, so that protein extraction was more difficult.

Since the sixth week gel gave the highest number of spots, the database for *B. alboglabra* leaves was established by using the protein pattern at the sixth week, as shown in Figure 3. A total of 103 protein spots were detected by using Image Master Software analysis. Protein spots were cut, digested and analysed by LC/MS/MS for protein identification. Peptides were searched by using the protein database search tool which includes ProteinLynx, SWISS-PROT, NCBI and Mascot. Proteins identified in *B. alboglabra* leaves are listed as Table in Supplementary material.

The 2-DE protein patterns from leaves at the second to the eighth week were matched by using Image Master Software and compared with the protein map from the sixth week leaves. The proteins from the second to the eighth week were classified into different functional classes as shown in Figure 4.

Protein functions were identified by using UniProtKB, and classified into nine classes consisting of photosynthesis and photorespiration, amino acid metabolism, carbon-compound and carbohydrate metabolism, protein metabolism, stress response, cellular communication and signal transduction, glycolysis and gluconeogenesis, unidentified and other

Figure 1. The development and growth of B. alboglabra observed at one to eight weeks. A: 1 week; B: 2 weeks; C: 3 weeks; D: 4 weeks; E: 5 weeks; F: 6 weeks; G: 7 weeks; H: 8 weeks using the same scale. The scale referred is 1 inch.

unknown proteins. The unidentified proteins refer to proteins with other functions and unknown refer to proteins with unknown function.

From the graph, similar patterns were observed in all classes with the lowest number of protein spots in the second week and increasing gradually to the sixth week, which had the highest number of protein spots, presumably due to the metabolism needed to support growth and development of the plants. The number of protein spots decreased starting the seventh week, possibly due to the plants becoming thicker making protein harder to extract, as they reached senescent stage.

The photosynthesis and photorespiration class showed the highest number of proteins in the leaves, followed by unknown proteins, protein metabolism, glycolysis and gluconeogenesis, stress response, amino acid metabolism, carbon compound and carbohydrate metabolism, cellular communication and signal transduction and unclassified proteins.

Proteins involved in photosynthesis and photorespiration were mainly oxygen evolving proteins, photosystem proteins, ribulose bisphosphate carboxylase/oxygenase and rubisco activase. These proteins play vital roles in photosynthesis and photorespiration required to provide energy and substrates for other pathways.

The selected protein spots in *B. alboglabra* leaves from the second to the eighth week of protein expressions are shown in Figure 5. The five listed proteins were involved in photosynthesis and photorespiration, namely oxygen evolving enhancer protein 2 chloroplastic, photosystem II protein, ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit, ribulose-1,5 bisphosphate activase and rubisco activase. These proteins showed increasing patterns from second to the sixth weeks and gradually decrease from the seventh to the eighth week.

4. Discussion

The protein identification of *B. alboglabra* has not yet been established, and this study would provide a baseline for the assessment of plant stress, tolerance and changes during plant growth and development. The vegetable, like other higher plants, possesses many unique pathways which include photosynthesis to provide energy for its growth (Bryce, 2003). Plant proteomics has been intensely studied since the availability of the full 2-DE mapping of the reference plant, *Arabidopsis thaliana*. Since then, progress has been made towards identifying and cataloguing proteins from various plants which include rice, soybean, barley and maize corresponding to their developmental processes (Salekdeh *et al.* 2007). There is no reported proteomic study of *B. alboglabra*, so far. In this study, the protein database was established as a reference gel from sixth week *B. alboglabra* leaves.

In the time course study, the seeds were grown on soil supplied with organic and chemical fertilizers with no other treatment except for water and harvested from two to eight weeks. The plants were harvested in the morning to minimize the possibility of protein changes. The proteins from leaves

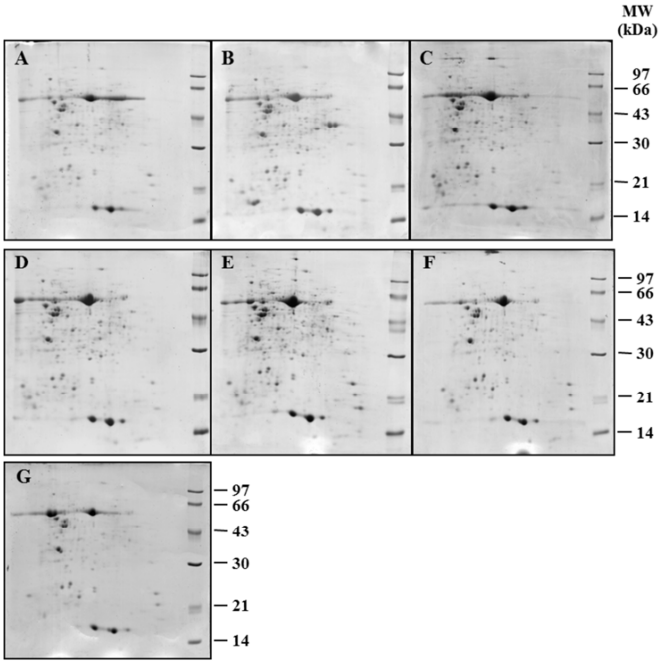


Figure 2. Two-dimensional PAGE of B. alboglabra (A-F) at different stages of growth. A: 2 weeks; B: 3 weeks; C: 4 weeks; D: 5 weeks; E: 6 weeks; F: 7 weeks; G: 8 weeks. Proteins (150 µg) were separated by using non-linear IEF strips of pH 3-10, followed by 12.5 % SDS PAGE and stained with Coomassie blue.

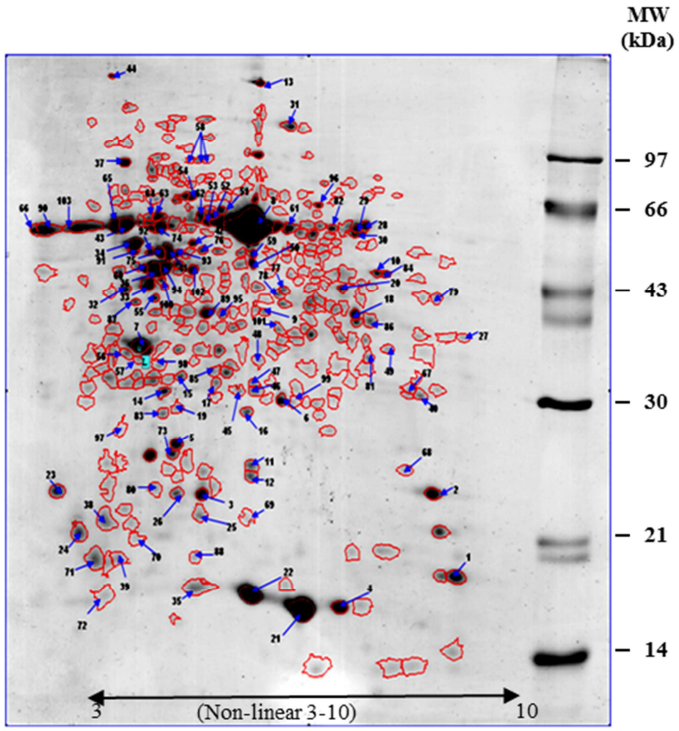


Figure 3. Protein map from B. alboglabra leaves at 6 weeks detected by using the Image Master Software. A total of 103 protein spots were identified from the database search performed with ProteinLynx screening SWISS-PROT, NCBI and the Mascot search tool available on the Matrix Science site screening.

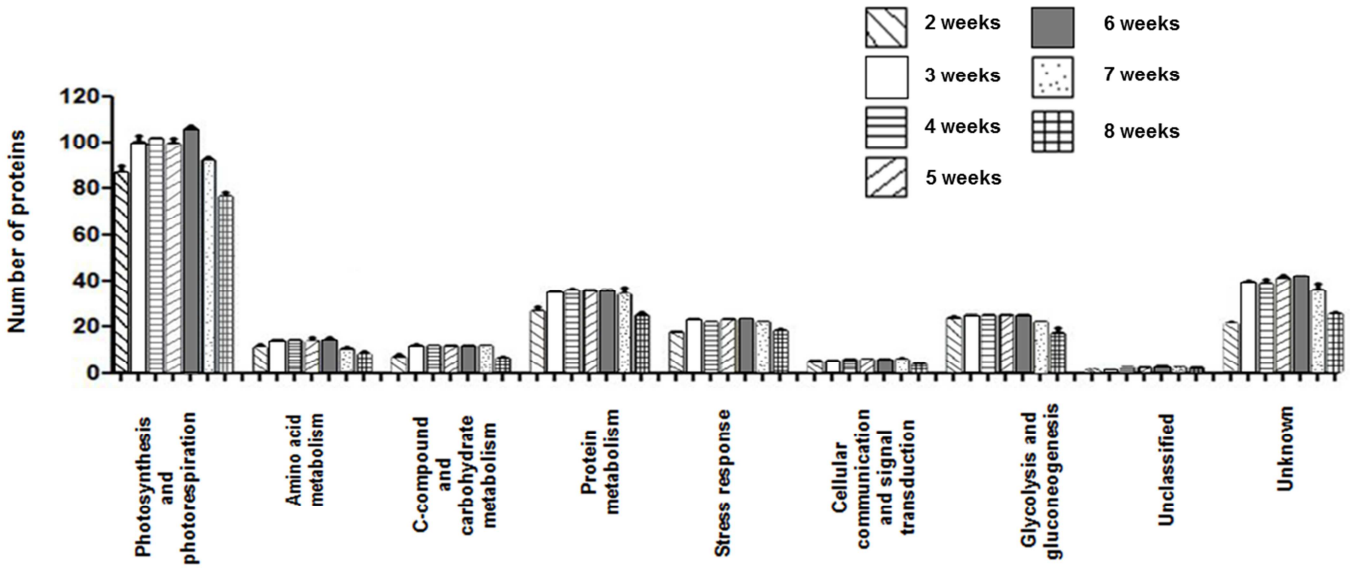


Figure 4. Protein spots of B. alboglabra leaves from two to eight weeks were grouped in nine functional classes, according to the main functions listed in UniProt (www.uniprot.org).

were identified and categorized into nine groups consisting of photosynthesis and photorespiration, amino acid metabolism, carbon-compound and carbohydrate metabolism, protein metabolism, stress response, cellular communication and signal transduction, glycolysis and gluconeogenesis, unidentified proteins and other proteins which are unclassified. The proteins showed almost similar increasing patterns from two to six weeks and began to decrease at seven weeks.

From the functional classification, photosynthesis and photorespiration proteins were dominantly found in the leaves. The proteins involved in photosynthesis and photorespiration were mainly from rubisco family, photosystem subunits oxygen evolving proteins and ATP synthase. During the development, the stems and leaves increased in size, and therefore plants require energy to promote biosynthesis and growth by the process of photosynthesis and photorespiration. The final products of photosynthesis are sucrose and starch which will be oxidized and enter pathways of proteins and cell wall synthesis (Bryce, 2003). Rubisco is the most abundant protein found in the leaves where large amounts of Rubisco are required to support photosynthesis (Good, 2003).

Oxidative stress has been proposed to play a developmental role in plants. Heat shock proteins and chaperonins are involved in stabilizing proteins folding in different kind of stimuli (Negri *et al*., 2008). Proteins involved in amino acid metabolism includes aminotransferases play roles in transporting nitrogen during development of cotyledons of germinating seeds to expand root, shoots and leaves (Lea, 2003).

Glycolysis and gluconeogenesis are the processes of breaking down and synthesizing sucrose or reducing sugars to provide energy and carbon skeletons for plant development, respectively (Bryce, 2003). The carbon metabolism of leaves is a light-independent process for carbon fixation to produce 3

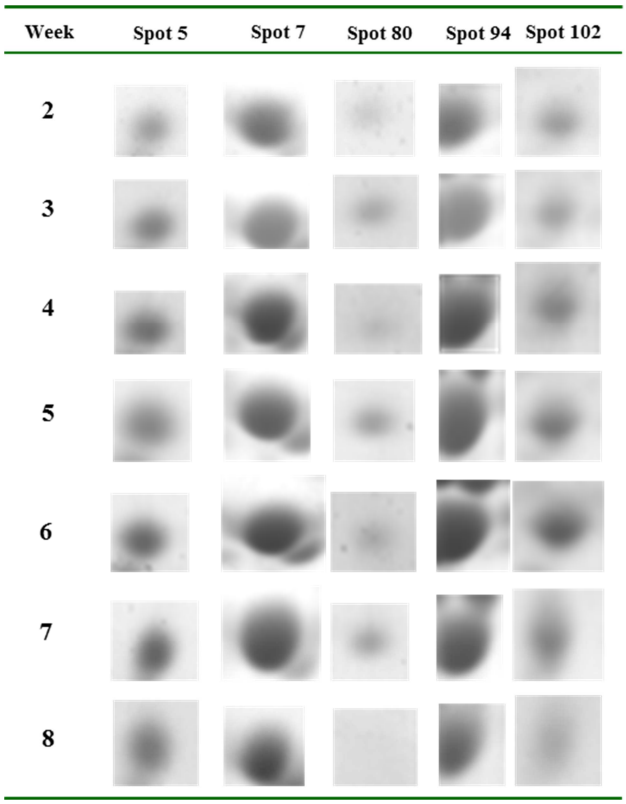


Figure 5. Five selected protein spots involved in photosynthesis and photorespiration from the second to the eighth week of B. alboglabra leaves. Spot 5 = Oxygen evolving enhancer protein 2 chloroplastic; Spot 7 = Oxygen evolving enhancer protein 1, photosystem II protein; Spot 80 = Ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit; Spot 94 = Ribulose bisphosphate carboxylase/ oxygenase activase; Spot 102= Rubisco activase

carbon sugars, unlike photosynthesis product of 6 carbon sugars. The 3 carbon sugars are converted to produce larger carbohydrates (Good, 2003). The skeleton of carbohydrates is used as the carbon source and energy such as cellulose to provide support to the plant the same way chitin does to insects. The major difference between carbohydrate metabolism in plants and animal is the synthesis of cell wall in plants which composed of mixture of complex polysaccharides (Smith, 2003).

Other proteins characterized were involved in protein synthesis and transcription grouped under protein metabolism, and proteins involved in signal transduction.

5. Concluding remarks

Proteomic studies of *B. alboglabra* have not yet been reported. Since this is the first proteomic study of *B. alboglabra*, database for the leaves had been established. The proteins were classified according to their functions as in photosynthesis and photorespiration, amino acid metabolism, carbon-compound and carbohydrate metabolism, protein metabolism, stress response, cellular communication and signal transduction, glycolysis and gluconeogenesis, unidentified proteins and other proteins. Further proteomic studies in *B. alboglabra* would provide more information for protein database in this plant species.

6. Supplementary material

The Table of identified proteins in this manuscript is online available in the web page of JIOMICS.

http://www.jiomics.com/index.php/jio/rt/suppFiles/25/0

Acknowledgements

We wish to thank Colombo Plan and Thai Government for the scholarship grant for Master studies. This investigation was supported by the Chulabhorn Research Institute.

References

1. J. H. Bryce, S. A. Hill, in: P. J. Lea (Eds). Plant biochemistry and molecular biology, John Wiley & Sons Ltd. West Sussex, 1993, pp. 1-26.
2. C. Chen, J. Exp. Bot. 50 (1999) 677-687.
3. S. A. Goff, D. Ricke, T. H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchison, C. Martin, F. Katagiri, B. M. Lange, T. Moughamer, Y. Xia, P. Budworth, J. Zhong, T. Miguel, U. Paszkowski, S. Zhang, M. Colbert, W. L. Sun, L. Chen, B. Cooper, S. Park, T. C. Wood, L. Mao, P. Quail, R. Wing, R. Dean, Y. Yu, A. Zharkikh, R. Shen, S. Sahasrabudhe, A. Thomas, R. Cannings, A. Gutin, D. Pruss, J. Reid, S. Tavtigian, J. Mitchell, G. Eldredge, T. Scholl, R. M. Miller, S. Bhatnagar, N. Adey, T. Rubano, N. Tusneem, R. Robinson, J. Feldhaus, T. Macalma, A. Oliphant, S. Briggs, Sci. 296 (2002) 92-100.
4. R. C. Leegood, in: P. J. Lea (Eds). Plant biochemistry and molecular biology: John Wiley & Sons Ltd. West Sussex, 1993, pp. 27-45.
5. L. Hennig, Trends Plant Sci. 12 (2007) 287-293.
6. M. Issarakraisila, Q. Ma, D. W. Turner, Scientia horticulturae. 111 (2007) 107-113.
7. N. N. V. Kav, S. Srivastava, L. Goonewardene, S. F. Blade, Annals of the Applied Biology. 145 (2005) 217-230.
8. G. X. La, P. Fang, Y. B. Teng, Y. J. Li, X, Y Lin, J. Zhejiang Univ. Sci. B. 10 (2009) 454-64.
9. J. H. Lum, K. L. Fung, P. Y. Cheung, M. S. Wong, C. H. Lee, F. S. Kwok, M. C. Leung, P. K. Hui, S. C. Lo, Proteomics. 2 (2002) 1123-30.
10. P. J. Lea, in: P. J. Lea, (Eds), Plant biochemistry and molecular biology: John Wiley & Sons Ltd. West Sussex, 1993. pp. 155-80.
11. L. Ma, D. Dolphin, Phytochemistry. 50 (1999) 195-202.
12. J. L. Micol, Curr. Opin. Plant. Biol. 12 (2009) 9-16.
13. B. P. Mooney, J. A. Miernyk, C. M. Greenlief, J. J Thelen, Physiologia Plantarum. 128 (2006) 237-250.
14. A. S. Negri, B. Prinsi, M. Rossoni, O. Failla, A. Scienza, M. Cocucci, L. Espen, BMC Genomics. 9 (2008) 378.
15. Y. Nozu, A. Tsugita, K. Kamijo, Proteomics. 6 (2006) 3665-3670.
16. C. J. Smith, in: P. J. Lea, (Eds). Plant biochemistry and molecular biology: John Wiley & Sons Ltd. West Sussex, 2003, pp. 73-111.
17. Thai vegetables: http://www.ku.ac.th/AgrInfo/fruit/veget/index.html
18. P. Talalay, J. W. Fahey, J. Nutr. 131 (2001) 3027S-3033S.
19. G. van Poppel, D. T. Verhoeven, H. Verhagen, R. A. Goldbohm, Adv. Exp. Med. Biol. 472 (1999) 159-68.
20. W. H. Vensel, C. K. Tanaka, N. Cai, J. H. Wong, B. B. Buchanan, W. J. Hurkman, Proteomics. 6 (2005) 1594-1611.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.42

Proteomic analysis of the mitochondria-enriched fraction from diabetic rat skeletal muscle

Edel Mullen[1], Kay Ohlendieck*[1].

[1]Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland.

Received: 27 September 2010 **Accepted:** 1 November 2010 **Available Online**: 9 November 2010

ABSTRACT

Mitochondrial dysfunction in muscle has been implicated to play a causative role or being an indirect consequence of insulin resistance in type-2 diabetes. In order to investigate potential diabetes-related alterations in the mitochondrial proteome of muscle, we have carried out a mass spectrometry-based proteomic analysis of the gastrocnemius muscle from normal versus diabetic Goto-Kakizaki rats. A generally perturbed protein expression pattern was observed in the mitochondria-enriched fraction from diabetic muscle. Various mitochondrial markers, including NADH dehydrogenase, cytochrome b-c1 complex and isocitrate dehydrogenase were reduced in diabetic preparations. Isoforms of pyruvate dehydrogenase and ATP synthase exhibited differential changes in their abundance. The altered protein expression levels of these key metabolic enzymes might trigger a diabetes-dependent decrease in mitochondrial oxidative phosphorylation levels. The proteomic findings presented here support the idea that mitochondrial abnormalities are involved in the molecular pathogenesis of type-2 diabetes and may be crucial for the development of insulin resistance.

Keywords: Diabetes; DIGE; Goto-Kakizaki Rat; Mitochondria; Muscle Proteomics.

1. Introduction

Besides being involved in intermediary metabolism, protein transport, cell cycle progression, calcium signaling and regulation of apoptosis, mitochondria represent the primary site for energy generation via oxidative phosphorylation [1]. Altered functioning or changed expression levels of mitochondrial proteins play a key role during development, natural aging and in various pathologies [2, 3]. Mass spectrometry-based subproteomics suggests itself as an ideal analytical method to determine global changes in the mitochondrial protein complement [4-6]. Proteomic screening of mitochondrial preparations suggest the existence of approximately 1,500 individual mitochondrial protein species, whereby considerable tissue-specific differences exist within the mitochondrial proteome [4]. The recent difference in-gel electrophoretic analysis of the mitochondria-enriched fraction from aged rat skeletal muscle has revealed drastic changes in many metabolic enzymes involved in oxidative phosphorylation [7]. The natural functional decline of the aged neuromuscular system is clearly associated with a shift to more oxidative-aerobic metabolism in a slower twitching fibre population [8]. In analogy, we have used here a similar proteomic approach to determine whether type-2 diabetes is associated with significant alterations in the mitochondrial proteome from skeletal muscle. We have focused our two-dimensional gel electrophoretic study on the pH 4 to 7 range since most soluble mitochondrial proteins can be separated in this acidic to neutral environment.

Impaired glucose homeostasis and diabetic complications affect millions of patients worldwide [9]. Insulin resistance impinges especially on contractile fibres, making abnormal glucose handling in skeletal muscle tissues a key issue for treating side effects of type-2 diabetes [10]. The etiology of diabetes mellitus appears to be influenced by both environmental and genetic factors, whereby obesity and a sedentary lifestyle have an extremely negative influence on disease progression [11]. Many diabetics suffer from glucotoxic side

***Corresponding author :** Dr. Kay Ohlendieck, Professor and Chair, Department of Biology, National University of Ireland Maynooth, Co. Kildare, Ireland; Tel. (353) (1) 708-3842, Fax. (353) (1) 708-3845; E-mail Address: kay.ohlendieck@nuim.ie.

effects and severely disturbed metabolic homeostasis [12]. Since muscle fibres are the most abundant cell type in the body and skeletal muscles represent the major organ for insulin-mediated whole body glucose disposal, insulin resistance in the musculature is one of the most striking pathophysiological features of type-2 diabetes [13]. In this respect, cellular dysfunction of muscle mitochondria is believed to play a crucial part in triggering insulin resistance [14-16]. It is not clear whether mitochondrial abnormalities are a secondary consequence of impaired hormonal signaling or the actual primary cause for a diminished glucose uptake into peripheral tissues [17]. However, independent of the exact initial role of mitochondria in diabetes, the long-term pathological consequences of disturbed mitochondrial bioenergetic functions are clearly a decline in crucial ATP-dependent cellular processes [16].

In order to determine potential diabetes-related changes in the mitochondrial subproteome from skeletal muscle, we have performed here a mass spectrometry-based proteomic analysis of the gastrocnemius muscle from normal versus diabetic Goto-Kakizaki (GK) rats. The non-obese GK animal model of type-2 diabetes is a spontaneously diabetic rat that shows clear signs of faulty insulin signaling within 4 weeks after birth [18]. GK rats exhibit increased levels of blood glucose, but no major changes in the concentration of non-fasting plasma insulin [19]. Diabetic GK skeletal muscles show: (i) an inhibition of insulin receptor autophosphorylation [20], (ii) impaired activities of insulin signaling intermediates [21], (iii) a diminished recruitment of the glucose transporter GLUT4 [22], (iv) membrane cytoskeletal defects in the dystrophin-dystroglycan complex [23], (v) a reduced percentage of oxidative fibres [24], (vi) a generally perturbed protein expression pattern [25], and (vii) abnormal mitochondrial functioning [26].

The proteomic analysis of the mitochondria-enriched fraction from normal versus diabetic skeletal muscles presented here has revealed a reduced expression of several mitochondrial enzymes, including NADH dehydrogenase, cytochrome b-c1 complex and isocitrate dehydrogenase. The changed abundance of these metabolic proteins might play a central role in the well-established decrease of mitochondrial oxidative phosphorylation in diabetic fibres. Hence, the proteomic results shown in this report have demonstrated that the mitochondrial subproteome is altered in type-2 diabetes and suggest that the changed density of key mitochondrial enzymes might aid in the pathological development of insulin resistance.

2. Material and methods

2.1 *Materials*

For the gel electrophoretic separation of muscle proteins, materials and analytical-grade chemicals were purchased from Amersham Biosciences/GE Healthcare, Little Chalfont, Buckinghamshire, UK (Imobiline IPG dry-strips, ampholytes, acetonitrile, Destreak agent, iodoacetamide, CyDye DIGE fluor minimal dyes Cy3 and Cy5), Biorad Laboratories, Hemel-Hempstead, Hertfordshire, UK (Laemmli-type buffer system; protein molecular mass markers) and National Diagnostics, Atlanta, GA, USA (ultrapure Protogel acrylamide stock solution). Protease inhibitors were obtained from Roche Diagnostics (Mannheim, Germany). Sequencing grade-modified trypsin was from Promega, Madison, WI, USA. Nitrocellulose transfer stacks were from Invitrogen (Carlsbad, CA, USA) and chemiluminescence substrate was purchased from Roche Diagnostics (Manheim, Germany). Primary antibodies ab66484 to mitofilin and ab14730 to the beta-subunit of mitochondrial ATP synthase were obtained from Abcam Ltd. (Cambridge, UK). Peroxidase-conjugated secondary antibodies were obtained from Chemicon International (Temecula, CA, USA).All other chemicals used were of analytical grade and purchased from Sigma Chemical Company, Dorset, UK.

2.2 *Preparation of mitochondria-enriched fraction from skeletal muscle*

Gastrocnemius muscle tissue (3g wet weight) from 9-week old normal versus age-matched diabetic GK rats was cut into small pieces and homogenised in 20 ml of buffer (220mM mannitol, 70mM sucrose, 20mM Hepes, pH 7.4), supplemented with 1 Roche PIC complete mini tablet per 10 ml of solution [7]. The resulting mixture was centrifuged at 1,100*g* for 5 min at 4ºC to remove cellular debris. The supernatant was retained. The pellet was re-suspended in 10 ml of homogenisation buffer and re-centrifuged as described above. Supernatants from the above separation steps were pooled and centrifuged at 7,000*g* for 15 min. The mitochondrial fraction was obtained as a pellet after repeated centrifugation steps at 20,000*g* for 15 min. The final mitochondria-enriched pellet was re-suspended in 2D lysis buffer for DIGE analysis. Samples were adjusted to a pH-value of 8.5. The protein content of individual samples was determined by the Bradford assay system [27].

2.3 *Fluorescent labelling of proteins*

Potential differences in the protein expression patterns of normal versus diabetic mitochondria were determined by difference in-gel electrophoresis (DIGE) using the CyDye DIGE fluor minimal dyes Cy3 and Cy5. DIGE dyes were reconstituted as a stock solution of 1mM in fresh dimethylformamide [9]. The stock solution was diluted to a working solution of 0.2mM prior to fluorescent labelling. Samples of 50 mg protein, representing each biological replicate, were minimally labelled with 200 pmols of Cy3 working solution. A pooled sample consisting of equal quantities of protein from all replicates used in the experiment were labelled at 200 pmol of Cy5 working solution to 50 mg of protein. The pooled fraction served as an internal standard. All samples were labelled at pH 8.5 with the appropriate amount of CyDye and after brief vortexing incubated on ice in the dark for 30 min. For the electrophoresis of individual gels, Cy3- and Cy5-labelled samples were mixed at an equal ratio. The

labeling reaction was terminated by the addition of 1µl of 10 mM lysine per 25 µg of muscle protein. Suspensions were briefly vortexed and then centrifuged at 12,000*g* for 10s and incubated on ice in the dark for 10 min. For gel electrophoretic separation, samples where loaded onto IPG strips with an equal volume of 2x sample buffer (7M urea, 2M thiourea, 65mM CHAPS, 2% ampholytes and 2% DTT) during rehydration [28].

2.4 *Two-dimensional gel electrophoresis*

The fluorescently labelled subproteomes from the mitochondria-enriched fraction of normal (n=4) versus diabetic (n=4) skeletal muscle tissues were separated in the first dimension by isoelectric focusing and in the second dimension by sodium dodecyl sulphate polyacrylamide gel electrophoresis, as previously optimized by our laboratory [25]. All biological replicates were analyzed individually. IPG strips were rehydrated in rehydration buffer (7M urea, 2M thiourea, 4% (w/v) CHAPS, 1.2% Destreak agent and 2% (v/v) pH 3-10 ampholytes) and 700 mg of protein sample for 12 hours. The first-dimension protein separation was conducted using the IPG strips on an Amersham IPGphor IEF system following the protocol previously described in detail [28]. First dimension strips were subjected to reduction and alkylation prior to second-dimension separation on 12.5% (w/v) slab gels. Gels were electrophoresed in an Amersham Ettan DALT-Twelve system [28].

2.5 *Image Acquisition and data analysis*

Fluorescently-labeled muscle proteins were visualised with the help of a Typhoon Trio variable mode imager from Amersham Biosciences/GE Healthcare (Little Chalfont, Bucks., UK). Two-dimensional gel images were analysed using Progensis SameSpots analysis software version 3.2.3 from Non Linear Dynamics (Newcastle upon Tyne, UK). All analytical DIGE gels were alligned to a reference gel. Following detection and filtering of spots, images were separated into groups (Normal rat muscle versus Goto-Kakizaki rat muscle) and analysed to determine significant changes in two-dimensional spot abundance. A hit list was generated of protein species that changed in density. An Anova score was included for each spot and any muscle proteins with an Anova score above 0.5 were excluded from further consideration. Any two-dimensional protein spot with a significant change in abundance was subsequently identified by mass spectrometry. Protein spots were excised from DIGE gels that had subsequently been stained with colloidal Coomassie Blue.

2.6 *Mass spectrometric identification of muscle proteins*

In order to identify individual muscle-associated protein species, peptide mixtures were analysed by mass spectrometry on a Model 6430 Ion Trap LC/MS apparatus from Agilent Technologies (Santa Clara, CA, USA). Excision, washing, destaining and treatment with trypsin were performed by a previously optimised method [7]. Following tryptic digestion, generated peptides were obtained by removing supernatants from digested gel plugs. Further recovery was achieved by adding 30% acetonitrile/0.2% trifluoricacid to the gel plugs for 10 min at 37 ^{0}C with gentle agitation. Peptide-containing supernatants were pooled and samples dried through vacuum centrifugation. Peptide mixtures were then re-suspended in 10µl mass spectrometry-grade water and 0.1% formic acid for identification by ion trap LC-MS analysis. Separation of peptides was performed with a nano flow Agilent 1200 series system, equipped with a Zorbax 300SB C18 5µm, 4 mm 40nl pre-column and a Zorbax 300SB C18 5µm, 43mm x 75µm analytical reversed- phase column using the HPLC-Chip technology. Mobile phases utilized were A: 0.1% formic acid, B: 50% acetonitrile and 0.1% formic acid. Samples (5µl) were loaded into the enrichment at a capillary flow rate set to 2 µl/min with a mix of A and B at a ratio 19:1. Tryptic peptides were eluted with a linear gradient of 10-90% solvent B over 2 µl/min with a constant nano pump flow of 0.6 ml/min. A 1 min post time of solvent A was used to remove sample carry over. The capillary voltage was set to 1700 V. The flow and the temperature of the drying gas were 4µl/ min and 300°C, respectively [7]. Database searches were carried out with Spectrum Mill Work Bench or Mascot MS/MS Ion search (Matrix Science, London, UK).

2.7 *Immunoblot analysis*

Standard one-dimensional immunoblotting was carried out in order to survey the mitochondrial isolation procedure and to validate expression changes in marker proteins as judged by DIGE analysis. Our optimized immunoblotting procedure for the immuno-decoration of muscle proteins has been previously described in detail [7]. Crude tissue extracts and the mitochondria-enriched fraction from normal versus diabetic skeletal muscle were electrophoretically separated with the help of a Mini-Protean-3 gel system (BioRad Laboratories, Hemel-Hempstead, Herts., UK). Gels were electrophoresed at 80V until the tracking dye ran off the end of the gel. For immunoblotting experiments, gels were transferred to an iblot transfer unit from Invitrogen (Carlsbad, CA, USA) for semi-dry blotting. Electrophoretic transfer was carried out for 6 min. Nitrocellulose membranes were blocked in a milk protein solution for 1h and then incubated overnight with gentle agitation with sufficiently diluted primary antibody, diluted in blocking solution containing 5% (w/v) fat-free milk powder in phosphate-buffered saline (0.9 % (w/v) NaCl, 50 mM sodium phosphate, pH 7.4). Following two washing steps with blocking solution for 10 min, sheets were incubated for 1h with secondary peroxidase-conjugated antibodies. Subsequently, nitrocellulose sheets were washed twice with blocking solution and then rinsed with phosphate-buffered saline. Immuno-decorated bands were visualized by the enhanced chemiluminescence method.

3. Results

Difference in-gel electrophoresis (DIGE) is a highly discriminating technique that can detect minute changes in the expression of soluble proteins. This makes DIGE analysis the

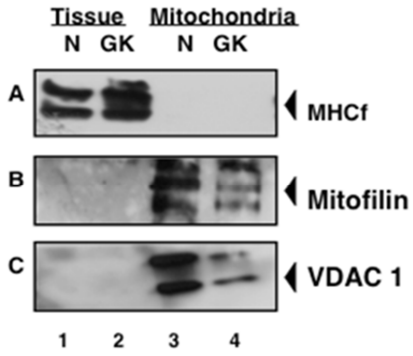


Figure 1. Immunoblot analysis of mitochondrial enrichment from normal and diabetic skeletal muscle. Shown are immunoblots of crude homogenates versus the mitochondria-enriched fraction from normal (N) versus Goto-Kakizaki (GK) rat muscle, labeled with antibodies to the fast MHCf isoform of myosin heavy cahin (A), the mitochondrial inner membrane marker mitofilin (B) and the mitochondrial outer membrane porin isoform VDAC1 of the voltage-dependent anion channel (C). Lanes 1 to 4 show tissue homogenates from normal and GK rat muscle and the mitochondrial fraction from normal and GK rat muscle, respectively.

method of choice for comparative biochemical studies of distinct protein complements [29]. Although two-dimensional gel electrophoresis underestimates the presence of certain classes of proteins, such as integral proteins and components with a very high molecular mass, the highly sensitive DIGE technique is ideal for the study of soluble proteins in the molecular range of 10 kDa to 200 kDa. We have applied here the DIGE method for the comparative analysis of the soluble mitochondrial subproteome from normal versus diabetic skeletal muscle tissue. As illustrated in Figure 1, the subcellular fractionation method applied here clearly resulted in the enrichment of mitochondrial markers. While the fast myosin heavy chain isoform MHCf drastically decreased in abundance during the isolation of the mitochondrial fraction (Figure 1A), marker proteins of the inner and outer mitochondrial membrane, mitofilin and the porin isoform VDAC1, increased in their density (Figure 1B, C). Thus, this differential centrifugation protocol removed the most abundant type of contractile muscle proteins from the mitochondrial preparation.

Following fluorescent labeling with CyDyes and the high-resolution two-dimensional gel electrophoretic separation of the mitochondria-enriched fraction, as shown in Figure 2, densitometric scanning revealed a changed abundance in 24 distinct protein spots. A reduced expression was found for 18 muscle-associated proteins and an increased density for 6 proteins in the mitochondria-enriched fraction. The first dimensional position of certain muscle-associated proteins, such as phosphoglycerate kinase, diacylglycerol kinase and phosphofructokinase, did not agree with the theoretical p*I*-values of these enzymes. However, the recent large-scale analysis of the aged mitochondrial proteome from rat skeletal muscle has shown that the mitochondrial organelle contains over a thousand different protein species, whereby many proteins are represented by a large number of distinct two-dimensional spots in analytical gels [7]. Thus, a specific type of enzyme exists in multiple sub-species and this explains why theoretical pI-values do not always perfectly agree with actual p*I*-values following two-dimensional gel electrophoresis. It is actually one of the great advantages of proteomic technologies that they can differentiate between differently charged or sized sub-species of individual proteins.

The DIGE master gel shown in Figure 3 outlines the position of the 24 2D-spots with a changed abundance in diabetic preparations. Mass spectrometry identified the proteins with a decreased abundance as the glycolytic enzyme enolase (spots 1 and 7), the molecular chaperones Hsp60 (spot 3) and Hsp72 (spot 8), muscle glycogen phosphorylase (spot 4), the mitochondrial enzyme NADH dehydrogenase (spots 5 and 12), subunit-1 of the cytochrome b-c1 complex (spot 6), phosphoglycerate kinase (spot 9) muscle creatine kinase (spot 10), diacylglycerol kinase (spot 11), isocitrate dehydrogenase (spot 13), myosin binding protein H (spot 14), the mitochondrial Tu translation elongation factor (spot 16), the 78 kDa glucose-regulated protein (spot 17) and the fast MLC1-f isoform of myosin light chain (spot 18). Interestingly, differential changes of individual 2D spots representing pyruvate dehydrogenase (spots 15 and 22) and ATP synthase (spots 2, 19 and 24) were observed. Increased proteins were identified as triose phosphate isomerase (spot 20), phosphofructokinase (spot 21) and pyruvate kinase (spot 23). Table 1 correlates spot numbers on the DIGE master gel with the proteomic information on individual protein species as determined by mass spectrometry. The table lists the protein name, accession number, predicted isoelectric point, predicted molecular mass, number of matched peptides, Mascot score, the percent sequence coverage and fold-change. The main subcellular

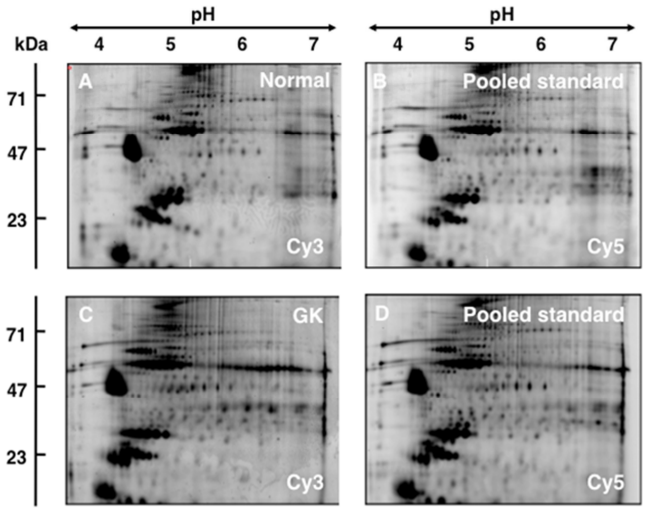


Figure 2. Comparative two-dimensional gel electrophoretic analysis of normal versus diabetic rat skeletal muscle. Shown is the 2D-DIGE analysis of the mitochondria-enriched fraction, using a pH 4-7 range in the first dimension. Individual panels represent Cy3-labelled gels of mitochondrial protein fractions from normal (A) and diabetic GK (C) skeletal muscle, as well as Cy5-labelled gels containing pooled standards (B, D). The pH-values of the first dimension gel system and molecular mass standards (in kDa) of the second dimension are indicated on the top and on the left of the panels, respectively.

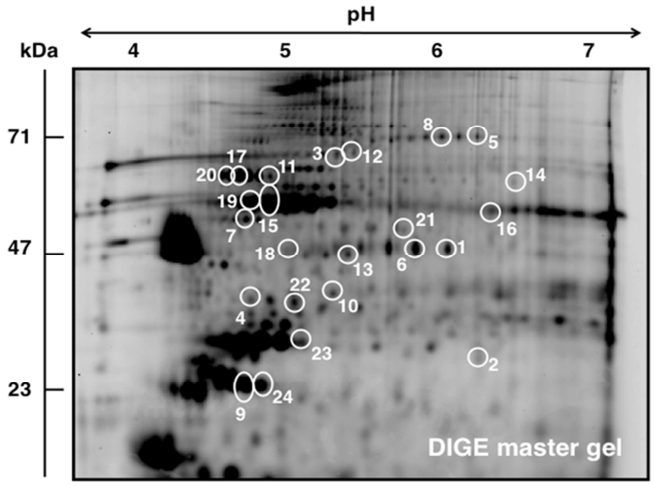


Figure 3. Two-dimensional DIGE reference gel of diabetic rat skeletal muscle. Shown is a DIGE reference gel of the mitochondria-enriched fraction from rat skeletal muscle, used for the mass spectrometric identification of proteins with a differential expression profile. The pH-values of the first dimension gel system and molecular mass standards (in kDa) of the second dimension are indicated on the top and on the left of the panel, respectively. Identified muscle proteins are marked by circles and are numbered 1 to 24. See Table 1 for a detailed listing of proteins that exhibited a diabetes-associated change in their abundance.

localizations of the identified proteins are predicted to be mitochondria (ATP synthase, NADH dehydrogenase, cytochrome b-c1 complex, isocitrate dehydrogenase, pyruvate dehydrogenase, F1-ATPase, 78kDa glucose-regulated protein, creatine kinase, Tu translation elongation factor, Hsp60), the cytosol (enolase, Hsp72, glycogen phosphorylase, phosphoglycerate kinase, diacylglycerol kinase, triose phosphate isomerase, phosphofructokinase, pyruvate kinase) and the contractile apparatus (myosin binding protein H, myosin light chain MLC-1f).

In order to verify DIGE-determined changes in marker proteins, immunoblotting was carried out. As shown in Figure 4, the molecular chaperone Hsp60 was clearly found to be reduced in GK preparations (Figure 4A), which agrees with the proteomic finding of a reduction in spot no. 3 (Figure 3; Table 1). Since the DIGE analysis of mitochondrial ATP synthase had revealed differential effects of diabetes on different sub-species of this mitochondrial enzyme, it was of interest to evaluate its molecular fate by immunoblotting. As illustrated in Figure 4B, the beta-subunit of ATP synthase was found to be increased in abundance. This agrees with the proteomic identification of increased spot no. 24 (Figure 3; Table 1). The statistical evaluation of the comparative immunoblot analysis of Hsp60 and mitochondrial ATP synthase is shown in Figure 4C, D and demonstrates the significance of the altered expression changes of these two proteins.

4. Discussion

Abnormal mitochondrial functioning has been implicated to play a central role in the molecular pathogenesis of insulin resistance and contractile weakness in diabetic skeletal muscle tissues [17]. The proteomic DIGE analysis of the mitochondria-enriched fraction presented here confirms a disturbed protein expression pattern in the mitochondrial subproteome from GK gastrocnemius muscle. Our previous proteomic survey of crude total extracts from diabetic GK muscle had identified moderate differential expression patterns in 21 protein species. The diabetic phenotype seems to be associated with a generally altered composition of the muscle protein complement, affecting especially glucose, fatty acid, nucleotide and amino acid metabolism, as well as the contractile apparatus, the cellular stress response, the antioxidant defense system and detoxification mechanisms [25]. With respect to changes in mitochondria, a reduction in NADH dehydrogenase, cytochrome b-c1 complex and isocitrate dehydrogenase agrees with the previously reported decreased oxidative phosphorylation in type-2 diabetes [14-16].

The lower concentration of the molecular chaperones Hsp60 and Hsp72 indicates an impaired cellular stress response in GK muscle tissue, which might weaken the defense mechanisms of metabolically challenged skeletal muscles. Reduced levels of muscle glycogen phosphorylase could have a negative effect on the proper utilization of stored glycogen and a lower concentration of creatine kinase may negatively affect the creatine phosphate shuttle. Both reductions in key metabolic components probably worsen the bioenergetic status of diabetic fibres and explain the contractile weakness in certain skeletal muscles from patients with type-2 diabetes [30, 31]. Interestingly, glycolytic marker enzymes were differentially affected in GK muscle. The increased density of triose phosphate isomerase, phosphofructokinase and pyruvate kinase would suggest a higher glycolytic flux rate in tissues with a reduced mitochondrial content. Since phosphofructokinase and pyruvate kinase represent rate-limiting steps of glycolysis, this alteration might be interpreted as a glycolytic

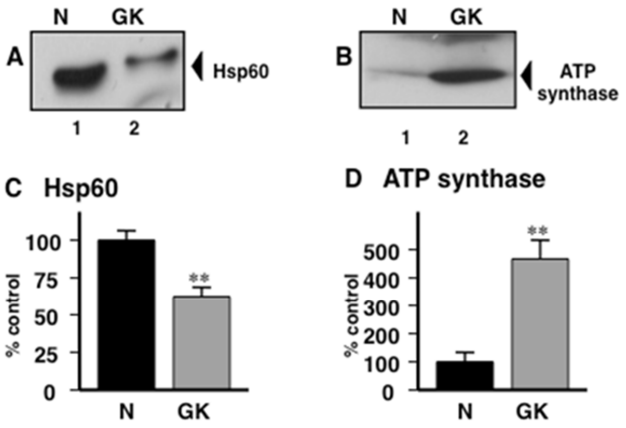


Figure 4. Immunoblot analysis of altered marker proteins in diabetic skeletal muscle. Shown are immunoblots of normal (lane 1) versus diabetic Goto-Kakizaki (lane 2) mitochondria-enriched preparations from skeletal muscle. Immuno-decoration was carried out with antibodies to the molecular chaperone Hsp60 (A, C) and the mitochondrial enzyme ATP synthase (B, D). The comparative blotting was statistically evaluated using an unpaired Student's t-test (n=5; **p<0.01).

Table 1. List of identified proteins that exhibit a changed abundance in mitochondria-enriched fraction from diabetic skeletal muscle.

Spot no.	Protein name	Accession no.	Isoelectric point (p*I*)	Molecular mass (kDa)	Matched peptides	ANOVA (p value)	Mascot score	Coverage (%)	Fold change
1	Enolase	gi|126723393|	7.08	47,332	17	0.009	314	44	-4.5
2	Mitochondrial H-ATP synthase	gi|220904|	5.78	18,828	3	0.021	74	16	-3.3
3	Heat shock protein Hsp60	gi|56383|	5.11	61,101	5	0.043	100	12	-3.2
4	Muscle glycogen phosphorylase	gi|158138498|	6.65	97,749	10	0.032	112	13	-3.1
5	NADH dehydrogenase, FeS protein 8	gi|157821497|	5.87	24,419	6	0.031	56	25	-2.9
6	Cytochrome b-c1 complex, subunit 1	gi|51948476|	5.57	53,511	9	0.004	241	19	-2.7
7	Enolase	gi|126723393|	7.08	47,332	17	0.002	206	44	-2.6
8	Heat shock protein Hsp72	gi|347019|	5.43	71,116	12	0.011	113	17	-2.6
9	Phosphoglycerate kinase	gi|40254752|	8.02	44,916	9	0.003	96	29	-2.5
10	Muscle creatine kinase	gi|6978661|	6.57	43,224	7	0.015	111	17	-2.4
11	Diacylglycerol kinase, delta	gi|149037714|	8.09	126,993	2	0.008	50	1	-2.3
12	NADH dehydrogenase, FeS protein 1	gi|53850628|	5.65	80,348	2	0.032	44	4	-2.2
13	Isocitrate dehydrogenase	gi|16758446|	6.47	40,052	7	0.049	198	22	-2.2
14	Myosin binding protein H	gi|38303941|	6	53,044	7	0.048	177	23	-2.1
15	Pyruvate dehydrogenase	gi|56090293|	6.2	39,305	11	0.031	294	34	-2.1
16	Tu translation elongation factor, mitochondrial	gi|149067905|	7.65	44,023	2	0.020	74	10	-2
17	78kDa glucose-regulated protein	gi|25742763|	5.07	72,476	2	0.024	60	4	-2
18	Myosin light chain MLC-1f	gi|205485|	4.99	20,795	15	0.003	200	67	-2
19	F1-ATPase, beta subunit	gi|203033|	5.07	38,747	4	0.009	187	12	2
20	Triose phosphate isomerase	gi|38512111|	7.07	27,223	10	0.006	152	59	2.1
21	Phosphofructokinase	gi|62825891|	8.07	86,159	5	0.040	89	7	2.2
22	Pyruvate dehydrogenase	gi|56090293|	6.2	39,305	3	0.044	37	10	2.6
23	Pyruvate kinase	gi|16757994|	6.63	58,303	13	0.013	283	30	2.7
24	ATP synthase, mitochondrial F1 complex	gi|149029719|	6.43	23,434	8	0.004	94	53	3.1

shift that has previously been described in obese skeletal muscles [32]. On the other hand, the enzymes enolase and phosphoglycerate kinase are reduced in GK tissue, which might be associated with the multi-functionality of many glycolytic elements [33]. The enzyme diacylglycerol kinase catalyzes the conversion of diacylglycerol to phosphatidic acid and uses ATP as a source of the phosphate. Thus reduced ATP levels in diabetic muscle tissues due to a reduced content of functional mitochondria could affect this enzyme. Altered levels of the fast MLC1-f isoform of myosin light chain and myosin binding protein H indicate a certain degree of re-modeling of the contractile apparatus. However, since this subproteomic analysis was carried out with a mitochondria-enriched fraction, it is difficult to interpret changes in regulatory elements of the acto-myosin filaments. Compensatory mechanisms to counter-act the loss of mitochondrial functioning appear to be an increase in certain isoforms of pyruvate dehydrogenase and ATP synthase [17].

In conclusion, the novel candidate proteins with a changed expression level in the mitochondria-enriched fraction, as shown in this proteomic survey of the GK rat model of non-obeses type-2 diabetes, should be helpful for complementing the biomarker signature of diabetes mellitus [34-38]. Changes in the functioning and/or density of mitochondrial enzymes

may be useful for the identification of new therapeutic targets, the development of better diagnostic criteria, the improved monitoring of disease progression and the biomedical evaluation of experimental treatment regimes.

Acknowledgements

Research was supported by a project grant from the Irish Health Research Board (HRB-RP/2005/3) and an equipment grant from the Higher Education Authority (HEA-RERGS-07-NUIM).

References

1. H.M. McBride, M. Neuspiel, S. Wasiak, Curr. Biol. 16 (2006) R551-R560.
2. D.C. Chan, Cell 125 (2006) 1241-1252.
3. D.L. Johannsen, E. Ravussin, Curr. Opin. Pharmacol. 9 (2009) 780-786.
4. A. M. Distler, J. Kerner, C.L. Hoppel, Proteomics 8 92008) 4066-4082.
5. M. Elstner, C. Andreoli, T. Klopstock, T. Meitinger, H. Prokisch, Methods Enzymol. 457 (2009) 3-20.
6. C. Ruiz-Romero, F.J. Blanco, Mol. BioSyst. 5 (2009) 1130-1142.
7. K. O'Connell, K. Ohlendieck, Proteomics 9 (2009) 5509-5524.
8. P. Doran, P. Donoghue, K. O'Connell, J. Gannon, K. Ohlendieck, Proteomics 9 (2009) 989-1003.
9. H. King, R.E. Aubert, W.H. Herman, Diabetes Care 2 (1998) 1414-1431.
10. E. Phielix, M. Mensink M, Physiol. Behav. 94 (2008) 252-258.
11. M.T.T. Luna, Arch. Med. Res. 36 (2005) 210-222.
12. A.J. Scheen, Acta Clin. Belg. 58 (2003) 335-341.
13. K.F. Petersen, G.I. Shulman, Am. J. Cardiol. 90 (2002) 11G-18G.
14. B.B. Lowell, G.I. Shulman, Science 307 (2005) 384-387.
15. Zorzano A, Liesa M, Palacín M., Arch Physiol Biochem. 115 (2009) 1-12.
16. I. Pagel-Langenickel, J. Bao, L. Pang, M.N. Sack, Endocr. Rev. 31 (2010) 25-51.
17. J.F. Dumas, G. Simard, M. Flamment, P.H. Duclzeau, P. Ritz, Diabetes Metab. 35 (2009) 159-167.
18. A. Kitahara, T. Toyota, M. Kakizaki, Y. Goto Y, Tohoku J. Exp. Med. 126 (1978) 7-11.
19. S.M. Abdel-Halim, A. Guenifi, H. Luthman, V. Grill, S. Efendic, S.G. Ostenson, Diabetes 43 (1994) 281-288.
20. S.S. Dadke, H.C. Li, A.B. Kusari, N. Begum, J. Kusari, Biochem. Biophys. Res. Commun. 274 (2000) 583-589.
21. T.L. Steiler, D. Galuska, Y. Leng, A.V. Chibalin, M. Gilbert, J.R. Zierath, Endocrinology 144 (2003) 5259-5267.
22. C. Mulvey, E. Harno, A. Keenan, K. Ohlendieck, Eur. J. Cell Biol. 84 (2005) 867-883.
23. C. Mulvey, E. Mullen, K. Ohlendieck K, Mol. Med. Rep. 1 (2008) 297-306.
24. K. Yasuda, W. Nishikawa, N. Iwanaka, E. Nakamura, Y. Seino, K. Tsuda, A. Ishihara, Clin. Exp. Pharmacol. Physiol. 29 (2002) 1001-1008.
25. E. Mullen, K. Ohlendieck, Int. J. Mol. Med. 25 (2010) 445-458.
26. W. Shen, J. Hao, C. Tian, J. Ren, L. Yang, X. Li, C. Luo, C.W. Cotma, J. Liu, Plos ONE 3 (2008) e2328.
27. M.M. Bradford, Anal. Biochem. 72 (1976) 248-254.
28. P. Doran, K. O'Connell, J. Gannon, M. Kavanagh, K. Ohlendieck, Proteomics 8 (2008) 364-377.
29. S. Viswanathan, M. Unlu, J.S. Minden, Nat. Protoc. 1 (2006) 1351-1358.
30. H. Andersen, S. Nielsen, C.E. Mogensen, J. Jakobsen, Diabetes 53 (2004) 1543-1548.
31. S.W. Park, B.H. Goodpaster, E.S. Strotmeyer, N. de Rekeneire, T.B. Harris, A.V. Schwartz, F.A. Tylavsky, A.B. Newman, Diabetes 55 (2006) 1813-1818.
32. D.S. Hittel, Y. Hathout, E.P. Hoffman, J.A. Houmard, Diabetes 54 (2005) 1283-1288.
33. K. Ohlendieck, Biochim. Biophys. Acta 1804 (2010) 2089-2101.
34. M. Korc, Mol. Cell. Proteomics 2.6 92003) 399-404.
35. K. Hojlund, K. Wrzesinski, P.M. Larsen, S.J. Fey, P. Roepstorff, A. Handberg, F. Dela, J. Vinten, J.G. McCormack, C. Reynet, H. Beck-Nielsen, J. Biol. Chem. 278 (2003) 10436-10442.
36. F.B. Stentz, A.E. Kitabchi, Genom. Proteom. Bioinformatics 5 (2007) 216-235.
37. T. Sundsten, H. Ortsaeter, Mol. Cell. Endocrinol. 297 (2009) 93-103, 2009.
38. H Hwang, B.P. Bowen, N. Lefort, C.R. Flynn, E.A. De Filippis, C. Roberts, C.C. Smoke, C. Meyer, K. Hojlund, Z. Yi, L.J. Mandarino, Diabetes 59 (2010) 33-42.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ? PR@AJC | DOI: 10.5584/jiomics.v1i1.32

Molecular Modeling of Cathepsin B protein in different Leishmania strains

Pawan Kumar Jayaswal[1,+], Mukta Rani[1], Chandra Prakash Yadav[1], Manas Ranjan Dikhit[1], Ganesh Chandra Sahoo*[1], Pradeep Das[1].

[1]Rajendra Memorial Research Institute of Medical Sciences, Agam Kuan, Patna, India- 80007; [+]Present address of the author is National Research Centre on Plant Biotechnology Genoinformatics, LBS Building, IARI, Pusa, New Delhi-110012.

Received: 13 August 2010 **Accepted:** 12 November 2010 **Available Online**: 20 November 2010

ABSTRACT

Cathepsin B like cysteine proteases representing a major component of the lysosomal proteolytic repertoire plays an important role in intracellular protein degradation. Comparative models of cathepsin B (CatB) protein of six different *Leishmania* strains were developed using MODELLER. The modeled three-dimensional (3-D) structure has the correct stereochemistry as gauged from the Ramachandran plot and good 3-D structure compatibility as assessed by PROCHECK and the DOPE score (DS2.1, Accelrys). The modeled proteins were energy minimized and validated using standard dynamic cascade protocol (DS 2.1). Seven different disulfide bonding sites are predicted in CatB protein of *Leishmania*. Two domains were identified and different motifs are present in catB protein of *Leishmania* like aspargine glycosylation sites, protein kinase phosphorylation sites, protein kinase C activation sites and N-myristoylation sites. Considering that cathepsin B is essential for survival of *Leishmania*, including for virulence to the mammalian host, it may be viewed as an attractive drug target.

Keywords: Cathepsin B; Homology Modeling; Leishmaniasis; Simulation; CatB; Cysteine protease.

1. Introduction

Leishmaniasis is a complex parasitic diseases caused by at least 17 different species of the protozoan parasite *Leishmania* [1]. It is transmitted by the bite of *Phlebotomine* sand flies; *Leishmania* infects approximately 12 million people and is commonly endemic in tropical and subtropical regions of America, Africa, and the Indian subcontinent, as well as in the subtropics of Southeast Asia [2]. *Leishmania* species are diploid eukaryotes and are obligate intracellular protozoa that reside in macrophages of their mammalian hosts [3]. *Leishmania donovani* and *major* are the causative agents of old world Visceral and Cutaneous Leishmaniasis (VL and CL) respectively. Among the variety of disease manifestations, VL is a systemic disease and *Leishmania donovani* complex is fatal and is a serious health problem in many tropical and subtropical countries [4] whereas CL, caused by species such as *Leishmania major, Leishmania mexicana, Leishmania braziliensis, and Leishmania panamensis*, frequently self-cures within 3–18 months, leaving disfiguring scars [5]. In vertebrate hosts Leishmaniasis are transmitted by *Phlebotomine* sand flies, which acquire the pathogen by feeding on infected hosts and transmit them by regurgitating the parasite at the site of a subsequent blood meal [5]. While obtaining blood meal, sand flies inject saliva into the host's skin containing anti clotting, anti platelet and vasodilatory compounds that increase the hemorrhagic pool [6, 7]. *Leishmania* parasites contain high levels of cysteine proteases [8] represent a major component of the lysosomal proteolytic repertoire and play an important role in intracellular protein degradation [9] The lysosomal cysteine proteases, cathepsins B, H, L, S and C, are well characterized proteins with closely related amino acid sequences, belonging to the papain super family [10]. Lysosomal enzymes are synthesized in normal cellular processes as glycosylated higher molecular weight precursors, which, during their maturation, undergo several processing steps by limited proteolysis [11]. Cathepsin B maturation includes removal of the N-terminal propeptide, the C-terminal extension and a dipeptide between residues 47 and 50 (mature enzyme numbering). The product is an en-

***Corresponding author:** Dr. Ganesh Chandra Sahoo, Scientist & Head; BioMedical Informatics Center, RMRIMS, Agam Kuan, Patna -80007, India; Contact No. +91- 612 -2631565, Fax No. +91-612-2634379; E-mail Address: ganeshiitkgp@gmail.com

zymatically active molecule with two chains covalently cross-linked by a disulfide bridge [12, 13]. It exhibits both endopeptidase and exopeptidase activities [14] and shown that the exopeptidase activity is dependent on the presence of a specialized structural element, the occluding loop, which accepts the negative charge of the P2 carboxylate. CatB protein that belongs to the papain super family and shows high homology to cathepsins L, S and O, papain and actinidin, among others [10]. This super family encompasses a large number of cysteine proteases from sources as diverse as bacteria, plants and mammals [15]. Studies involving species of *Leishmania* such as *L. major* and *L. amazonensis* have been shown to induce the production of biologically active transforming growth factor Я (TGF-β) by macrophages upon infection [16]. *L. donovani* infection is known to induce the expression of a number of cytokine genes like TNF-Ю , GM-CSF, TGF-β, and IL-6 [17]. Application of anti-TGF-β antibodies arrested the development of lesions in mice, whereas treatment with TGF-β exacerbated the infection with *L. amazonensis*, *Trypanosoma cruzi*, and *Toxoplasma gondii* [18]. *Leishmania major, L. mexicana*, and *L. braziliensis* trigger the production of TGF-β and IL-10, which inhibit killing of intracellular organisms [19]. It was also shown that TGF-β plays a role in limiting IFN- γ production during the primary infection in mice [20]. IFN-γ, on the other hand, is known to induce the expression of inducible nitric-oxide synthase, the key effectors mechanism for the killing of *Leishmania* and other parasites within the mouse macrophages, both *in vitro* and *in vivo* [21]. TGF-β isoforms are synthesized as large biologically inactive precursors, which are proteolytically processed to yield mature and active homodimer. A variety of agents and treatments are known to activate latent TGF-β, like heat, acidic pH, plasmin, subtilisin-like endopeptidases, and cathepsins [22, 23, 24]. Hence, structural and functional analysis of catB protein of different *Leishmania* sp. is vital for further structure based ligand protein interaction study.

In this work, we are concerned with the determination of the 3-D structure of cathepsin B-like cysteine protease of six different strains of *Leishmania* (*L. donovani, L. infantum, L. major*, *L. mexicana, L. chagasi,* and *L. braziliensis*). These developed models will be helpful in the screening of several antileishmanial drug molecules and alkaloids in future. Screening of antileishmanial compounds (Ligand molecules) against cathepsin protein study is going on in Biomedical Informatics Centre (BIC) of RMRIMS with the different commercial software.

2. Material and methods

In this study various three dimensional structural models of the cathepsin B protein of six different *Leishmania* strains were generated. The models were validated by Ramachandran plots of PROCHECK and DOPE scores of Discovery Studio software v 2.1. The models of cathepsin B protein were further tested for insilico docking study to know the presence of any interaction between the ligand and Cat B protein. Ligand protein interaction of KMP-11 has already been reported earlier [25]. Various methods applied in this study are given below.

2.1 *Structural Modeling and Sequence Analysis of Cathepsin B Protein*

The amino acid sequence of cathepsin B of *L. donovani* (340 amino acids (aa), Genbank locus ID: AAG44365), *L. braziliensis* (340 aa, Uniprot ID: A4HH90), *L. infantum* (340 aa, Uniprot ID: A414D6), *L. chagasi* (340 aa, Uniprot ID: Q9GQN7), *L. major* (340 aa, Uniprot ID: Q4FXX7), and *L. mexicana* (340 aa, Uniprot ID: Q25319) was downloaded through NCBI & EMBL website for structural modeling. Multiple alignments of the related sequences were performed using Clustal W program accessible through the European Bioinformatics Institute [26] (http://www.ebi.ac.uk/Tools/clustalw2/index.html). No X-ray crystallographic or NMR structure of Cathepsin B protein of any *Leishmanial* strains has yet been determined. Tertiary structures of cathepsin B protein of six different *Leishmanial* strains were modeled on the basis of template pdb id: 3PBH & 1MIR using MODELER protocol of Discovery Studio 2.1. Structure of LPG2 protein of different *Leishmania* strains has already been reported earlier [27]. Structure validation was performed using Verify protein (DOPE) scores, WHATIF and, molecular modeling tools of Discovery studio. Cathepsin B protein of six different *Leishmania* strains and their two different template homologs and their PBB ID: 1MIR and 3PBH having their tertiary structures i.e. β sheets and α- helices are predicted through of Discovery Studio 2.1.

2.2 *Simulation of Cathepsin B protein*

Model of Cathepsin B protein of six different strains of *Leishmania* were further processed by applying CHARMM force field. Potential energy of a specified structure is evaluated by using calculate energy protocol of DS2.1. The calculate energy protocol can be used to compare the relative stability of different configurations of the same structure; or as a prelude to lengthy simulations to confirm the availability of appropriate force field parameters. The CHARMM molecular simulation package uses the CHARMM force field to model the energetic, forces and dynamics of biological molecules using the classical method of integrating Newton's equations of motion [28]. Energy minimization of al six different 3-D modeled protein structures are done with the help of standard dynamics cascade protocol of DS 2.1[29] which performs the following steps: minimization with steepest descent method, minimization with conjugate gradient, dynamics with heating, equilibration dynamics, production dynamics. The minimization protocol minimizes the energy of a structure through geometry optimization. The dynamics (heating or cooling) protocol allows controlling the temperature of a system when performing a molecular dynamics simulation. For the simulation cascade following parameter are used: steepest descents minimization (500steps, RMS gradient 0.1) in first minimization step & in second steepest Descents minimization (500 steps, RMS gradient 0.0001), heating (2000

steps , initial temperature 50K, final temperature 300K), equilibration (120 ps, 1fs time step, coordinates saved every 1000 steps) and Production (120 ps, 1fs time step, 300 K, NVT ensemble, non bond cutoff 14A, switching function applied between 10 and 12A, coordinates saved every 1000 steps).

2.3 *Function Assignment of Cathepsin B Protein by SVM*

To know the novel functions of Cathepsin B protein of all six different *Leishmania* strains were searched through BIDD server (http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi) [30]. The web-based software, SVMProt, support vector machine (SVM) classifies a protein into functional families from its primary sequence based on physico-chemical properties of amino acids. Novel protein function assignment of different proteins of SARS virus and Japanese encephalitis virus has already been reported by using this server [31].

2.4 *Predict Protein Server*

Predict Protein provides PROSITE sequence motifs, low complexity regions (SEG), nuclear localization signals, regions lacking regular structure (NORS) and predictions of secondary structure, solvent accessibility, globular regions, transmembrane helices, coiled-coil regions, structural switch regions, disulfide-bonds, sub-cellular localization, and functional annotations [32,33,34,35].

3. Results and Discussion

Structure, function and ligand binding site analysis of cathepsin B protein will lead to identification of novel targets for design of suitable lead compounds inhibiting the specific functions of *L. donovani, L. infantum, L. major, L. mexicana, L. chagasi,* and *L. braziliensis.*

3.1 *Structural Modeling and Sequence Analysis of Cathepsin B Protein*

Structural model of all six strains (*L. braziliensis, L. infantum, L. chagasi, L. major, L. mexicana, L. donovani*) of cathepsin B protein of *Leishmania* is modeled by MODELER on the basis of different three dimensional co-ordinates of two crystal structures (templates) of proteins namely PDB ID: 3PBH & 1MIR taken for this study from RCSB PDB are same for all six strains of *Leishmania* shown in Table 1. The PDB ID: 3PBH and 1MIR are selected as a template on the basis of BLAST result. Each strain of catB protein of *Leishmania* shows that the different identity with template protein. *L. donovani, L. braziliensis and L. infantum* sequences are having 38% sequence similarity with template 3PBH and 1MIR while *L. major and L. mexicana* is having 39% sequence similarity with 3PBH and 38% with 1MIR protein. *L. chagasi* shows only 37 % sequence similarity with 1MIR and 38% with 3PBH is shown in Table 1.

Crystallographic studies have demonstrated the structural features of template proteins i.e. 3PBH & 1MIR of human liver cathepsin B protein and rat procathepsin B respectively at resolution 2.8A° and 2.15A° respectively. The three dimensional structure of human procathepsin B (PDB ID: 3PBH) revels that the propeptide folds on the catB surface [36], shielding the enzyme active site from exposure to solvent. The structure of the enzymatically active domains is virtually identical to that of the native enzyme [37]. The three dimensional coordinates (PDB ID: 1MIR) represents cysteine protease of the papain super family which is synthesized as inactive precursors with a 60-110 residues at the N-terminal pro segment. The propeptide are potent inhibitors of their parent protease [38]. On the basis of above template studies it is hypothesized that the pro region lies in between the Tyr27 – Gln95 amino acids sequence of the different *Leishmania* strains i.e. *L. donovani, L. chagasi, L. infantum, L.*

Table 1. Relative Data of six different strains of Cathepsin B proteins of *Leishmania* as well as their templates (3PBH and 1MIR) shows the highest verify score value than the expected score value. Identity of template protein is varied between 37-39%. Abbreviations: LEIDO: *L. donovani,* LEIBR: *L. braziliensis,* LEIIN: *L. infantum*, LEICH: *L. chagasi* LEIME: *L. major,* LEIME: *L. mexicana.*

S.No.	Name of Target Sequence	Template protein			Model Name	Dope Score	Verify Score	Expected High Score	Expected Low Score
		Name	Length	Identity (%)					
1	AAG44365	3PBH	317	38	3PBH.B99990001	-33420.4	150.65	139.723	62.8754
	LEIDO	1MIR	313	38					
2	A4HH90 LEIBR	3PBH	317	38	1MIR.B99990002	-32069.9	146.29	139.723	62.8757
		1MIR	313	38					
3	A414D6	3PBH	317	38	1MIR.B99990001	-32335.4	153.31	139.723	62.8757
	LEIIN	1MIR	313	38					
4	Q9GQN7	3PBH	317	38	3PBH.B99990003	-33420.4	154.79	139.723	62.8757
	LEICH	1MIR	313	37					
5	QLFXX7	3PBH	317	39	3PBH.B99990003	-33420.4	154.79	139.723	62.8757
	LEIMA	1MIR	313	38					
6	Q25319	3PBH	317	39	3PBH.B99990003	-33420.4	154.79	139.723	62.8757
	LEIME	1MIR	313	38					

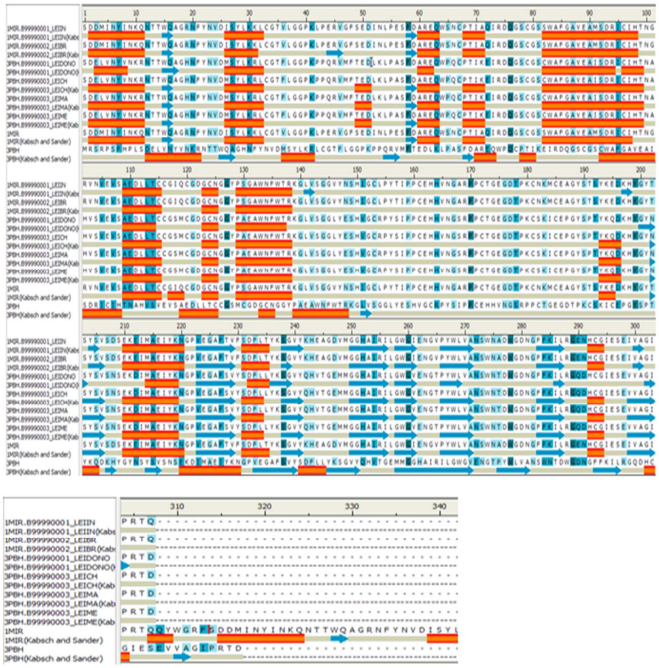


Figure 1. Comparative analysis of β-sheet and α-helix of protein PDB ID: 3PBH (3-D structure of human procathepsin B) and 1MIR (3-D coordinates of cysteine protease of papain super_ family) with cathepsin B protein of six different strains of Leishmania (DS 2.1). In figure1, Red color shows helical structure and blue arrows are the β-sheets (This Figure is available in the supplementary material).

major and *L. mexicana* but in case of *L. braziliensis* Gln95 amino acid residue is replaced by Ala95 residues of the protein sequence. After the cluster analysis of amino acid sequence of catB protein of six *Leishmania* strains and two templates it is known that the mature sequence begins at the N-terminal Met1 and ends at C-terminal Glu-340 amino acids. Secondary structure of catB protein (α -helices and β -sheets) are represented in Figure 1, 340 amino acid sequences of catB protein of six *Leishmania* strains and two template PDB ID: 3PBH and 1MIR in the modeled catB protein of *Leishmania,* 7-9 α helices and 6-8 β sheets have been observed.

Multiple alignment of amino acid sequences of cathepsin B protein of different *Leishmania* strains show that there is much identity (66-100%) among each other, catB protein of different *Leishmania* strains show 35- 45% identity with catB protein of human. Cathepsin B of *L. infantum* and *L. donovani* are identical (100%) to each other, hence demographic separation do not have any impact on protein structure at these two strains. Cathepsin B Protein of six different strains of *Leishmania* and three different strains of *Homo sapiens* i.e. B3KQRS, P07858 and B4DMY4 have been aligned through multiple sequence alignment by using Clustal W. It is learnt from multiple alignments that six strains of *Leishmania* are similar to each other and dissimilar with three different strains in human which is shown in Table 2. From cladogram, catB protein of *L. braziliensis* and other *Leishmania* strains are forming a cluster different from other cysteine proteases of human. Three strains of human cysteine proteins are found to be far from other *Leishmania* strains shown in Figure 2.

Developed 3- D models of all six *Leishmania* strains is verified with the help of Verify protein (MODELER) score protocol of DS2.1 and their score is higher than the expected high score. The DOPE (Discrete Optimized Protein Energy) score (-33420.4) is same for 3-D structure models of catB protein of *L. mexicana, L. major, L. chagasi* and *L. donovani* and is different i.e. (-32335.4) from other two modeled structures that of *L. braziliensis* and *L. infantum* is shown in Table1.

Cathepsin B protein of structural models of six *Leishmania* strains were validated by verify protein (MODELER) score of DS 2.1 (Accelrys), WHATIF and PROCHECK. In the Ramachandran plots (Procheck) show that 65-88.1% amino acid residues belong to core region, 9-30% residue in allowed region, 0.4 - 4.4% are in generously allowed regions and 0.4–1.2% in disallowed region is reported in Table 3. These amino acids which occur in invalid region of Ramachandran plot were further refined by side chain and loop refinement tools of DS2.1 (Accelrys) to get validate of the 3-D structure of cathepsin B protein. The best model of different *Leishmania* strains were screened by Verify protein (MODELER) score and the best was selected for further analysis. About 7 to 9 α-helices have been observed for catB protein of different models of various *Leishmania* strains, seven, eight, and nine helices have been observed in catB protein of modeled structure in *L. donovani, L. braziliensis, L. infantum* respectively. The modeled structures of catB protein of different *Leishmania* strains have shown close identity with each other, one modeled structure of *L. braziliensis* is given in Figure 3. The verify protein (MODELER) score of best predicted models of catB

Table 2. ClustalW results of multiple sequence alignment scores of Cathepsin B Protein of six different strains of Leishmania and three strains of human. Leishmania strains are closely associated with each other and far from the human catB protein sequence.

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	A4I4D6_L.IN_CPC_	340	2	Q9GQN7_L.CH	340	99
1	A4I4D6_L.IN_CPC_	340	3	Q4FXX7_L.MA_CPC_	340	91
1	A4I4D6_L.IN_CPC_	340	4	Q25319_L.ME	340	85
1	A4I4D6_L.IN_CPC_	340	5	A4HH90_L.BR_CPC_	340	68
1	A4I4D6_L.IN_CPC_	340	6	B3KQR5_H_cDNA_	339	35
1	A4I4D6_L.IN_CPC_	340	7	|P07858|CATB_H	339	35
1	A4I4D6_L.IN_CPC_	340	8	B4DMY4_H_cDNA_	245	42
1	A4I4D6_L.IN_CPC_	340	9	|gb|AAG44365.1|L.DO	340	100
2	Q9GQN7_L.CH	340	3	Q4FXX7_L.MA_CPC_	340	91
2	Q9GQN7_L.CH	340	4	Q25319_L.ME	340	85
2	Q9GQN7_L.CH	340	5	A4HH90_L.BR_CPC_	340	67
2	Q9GQN7_L.CH	340	6	B3KQR5_H_cDNA_	339	35
2	Q9GQN7_L.CH	340	7	|P07858|CATB_H	339	35
2	Q9GQN7_L.CH	340	8	B4DMY4_H_cDNA_	245	41
2	Q9GQN7_L.CH	340	9	|gb|AAG44365.1|L.DO	340	99
3	Q4FXX7_L.MA_CPC_	340	4	Q25319_L.ME	340	82
3	Q4FXX7_L.MA_CPC_	340	5	A4HH90_L.BR_CPC_	340	66
3	Q4FXX7_L.MA_CPC_	340	6	B3KQR5_H_cDNA_	339	36
3	Q4FXX7_L.MA_CPC_	340	7	|P07858|CATB_H	339	36
3	Q4FXX7_L.MA_CPC_	340	8	B4DMY4_H_cDNA_	245	42
3	Q4FXX7_L.MA_CPC_	340	9	|gb|AAG44365.1|L.DO	340	91
4	Q25319_L.ME	340	5	A4HH90_L.BR_CPC_	340	69
4	Q25319_L.ME	340	6	B3KQR5_H_cDNA_	339	36
4	Q25319_L.ME	340	7	|P07858|CATB_H	339	36
4	Q25319_L.ME	340	8	B4DMY4_H_cDNA_	245	42
4	Q25319_L.ME	340	9	|gb|AAG44365.1|L.DO	340	85
5	A4HH90_L.BR_CPC_	340	6	B3KQR5_H_cDNA_	339	40
5	A4HH90_L.BR_CPC_	340	7	|P07858|CATB_H	339	40
5	A4HH90_L.BR_CPC_	340	8	B4DMY4_H_cDNA_	245	44
5	A4HH90_L.BR_CPC_	340	9	|gb|AAG44365.1|L.DO	340	68
6	B3KQR5_H_cDNA_	339	7	|P07858|CATB_H	339	100
6	B3KQR5_H_cDNA_	339	8	B4DMY4_H_cDNA_	245	94
6	B3KQR5_H_cDNA_	339	9	|gb|AAG44365.1|L.DO	340	35
7	|P07858|CATB_H	339	8	B4DMY4_H_cDNA_	245	94
7	|P07858|CATB_H	339	9	|gb|AAG44365.1|L.DO	340	35
8	B4DMY4_H_cDNA_	245	9	|gb|AAG44365.1|L.DO	340	42

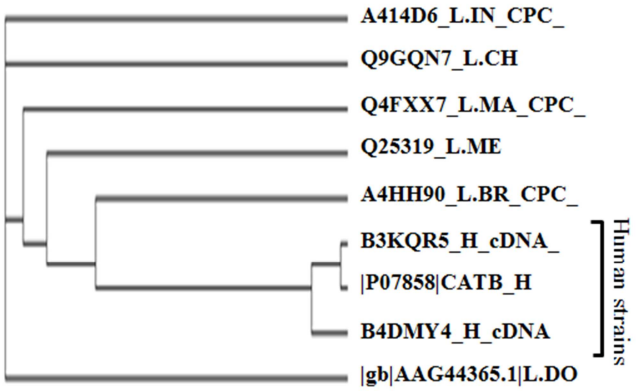


Figure 2. Phylogram showing phylogenetic relationship of Cathepsin B protein of six strains of Leishmania (L. infantum, L. chagasi, L. mexicana, L. braziliensis, L. major and L. donovani) with three different strains i.e. B3KQR5_H(cDNA), P07858,CATB_H and B4DMY4_H(cDNA) of cathepsin B protein in Homo sapiens.

Table 3. Referring to Ramachandran Plots of cathepsin B protein of six different strains of *Leishmania*. Abbreviations: Ldv: *L. donovani*, Lbrzl: *L. braziliensis*, Linf: *L. infantum*, Lch: *L. chagasi,* Lma: *L. major,* Lme: *L. mexicana.*

Residues	Number of Amino acids involved						Percentage of amino acids involved					
	Ldv	Lbrzl	Linf	Lch	Lma	Lme	Ldv	Lbrzl	Linf	Lch	Lma	Lme
Residues in most favoured regions [A,B,L]	164	223	220	220	220	220	65.1	88.1	87.0	87.3	87.3	87.3
Residues in additional allowed regions [a,b,l,p]	75	25	28	30	30	30	29.8	9.9	11.1	11.9	11.9	11.9
Residues in generously allowed regions [~a,~b,~l,~p]	11	3	2	1	1	1	4.4	1.2	0.8	0.4	0.4	0.4
Residues in disallowed regions	2	2	3	1	1	1	0.8	0.8	1.2	0.4	0.4	0.4
Number of non-glycine and non-proline residues	252	253	253	252	252	252	100% for all strains					
Number of end-residues (excl. Gly and Pro)	2	2	2	2	2	2						
Number of glycine residues (shown as triangles)	35	38	38	35	35	35						
Number of proline residues	18	14	14	18	18	18						
Total number of residues	307	307	307	307	307	307						

protein of various *Leishmania* strains Table 4, shows that highest scores (154.79) has been found in case of cathepsin protein model of *L. chagasi, L. major, L. mexicana* and lowest scores (146.29) has been found in case of *L. braziliensis*.

From Ramachandran plot, it is known that maximum residues in cathepsin B protein are responsible for construction of helices. It is found that the best model of cathepsin B protein of all these strains consisted of only one chain. In all the models 19 – 23 % is helical. The best models of catB protein of *L. chagasi, L. major, L. mexicana* are having highest number (eight) of helices where minimum three and maximum seventeen residues take part in formation of a helix. In all six strains of *Leishmania* 3_{10} helices have also been found where as in *L. donovani* maximum eighteen residues has been involved in forming a helix and helices in this strain accounts 19.2 % of all is shown in Table 4. CatB protein of six *Leishmania* strains that varies six to eight β sheets. Six β sheets are observed in *L. infantum* where as in L. *chagasi, L. major, L. mexicana* are having eight β sheets similarly seven β sheets are observed in *L. donovani and L. braziliensis.*

3.2 *Simulation of Cathepsin B protein*

Cathepsin B protein of six different strains of *Leishmania* are simulated by standard dynamic cascade protocol, in this process each simulation consists of 500 steps which is extended up to 5000 after that in each step 1000 increment has been given which was continue up to 10,000 steps of energy minimization. Each step can calculate the Van der Waals energy, CHARMM energy, potential energy and kinetic energy of the protein. Net partial charge and Net formal charge of catB protein of *L. donovani*, *L. chagasi, L.major, L. mexicana* are having -9 and that of other two strains are having -11 in *L. braziliensis and L. infantum*. Initial CHARMM energy of catB *L. chagasi, L. major, L. mexicana* was 14341 Kcal/mol and that of *L. donovani* was -19493.7 Kcal/mol and in *L. braziliensis* was -9705.34 Kcal/mol, after the minimization of energy up to 10,000 steps. CHARMM force field of cathepsin-B protein of each strain was changed and it varies between -16343.9 kcal/mol to -17110.1 kcal/mol. Van der Waals energy of cathepsin B of *L. donovani* changed from 2007.61 kcal/mol to -1732.12 kcal/mol, similar type of Van der Waals energy variation has been observed in other cathepsin B of five strains is shown in Table 5.

3.3 *Functional Assignment of Cathepsin B protein by SVM*

From the comparative analysis of cathepsin B protein of different *Leishmania* strains functional assignment shows that it belongs to transmembrane region protein. Cathepsin B protein of *L. donovani, L. chagasi, L. infantum* strains belongs to metal binding (65.4 %), manganese binding (62.2 %), copper binding (58.6 %) and magnesium binding (58.6 %) protein function families. Other protein functional families like hydrolases - acting on peptide bonds (peptidase) has been detected in *L. donovani*, *L. chagasi* and *L. infantum* (76.2%) and 85.4 % in *L. major* and 80.4 % in *L. mexicana*. Cathepsin B protein of *L. braziliensis, L. major and L. mexicana* belongs to metal binding functional motifs are 62.2 %, 73.8% and 71.3% respectively. Cathepsin B protein of *Leishmania major* has manganese binding (65.4 %), and 58.6% copper binding in *L. mexicana*. Calcium binding property has

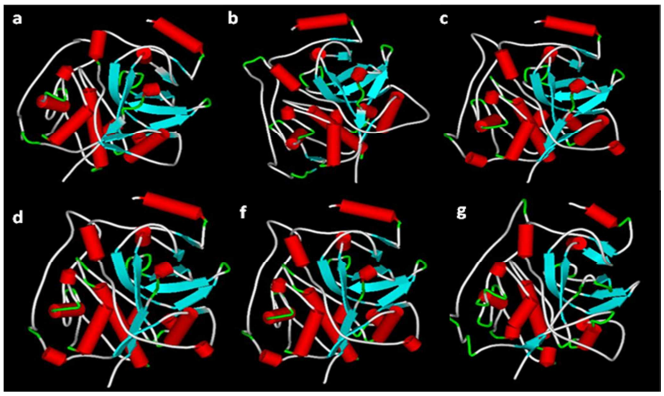


Figure 3. Ribbon representations of the homology model of Cathepsin B protein images of all six different leishmania strains using Discovery Studio 2.1 (Accelyrs) software (a) *L. braziliensis* (b) *L. infantum* (c) *L. chagasi* (d) *L. major* (e) *L. mexicana* and (f) *L. donovani.*

Table 4. Promotif search result summary and Profiles- 3D scores of modeled structure of cathepsin B proteins of all six strains of *Leishmania*.

Model Features Strain Names	No. and % of alpha helices	No. and % of 3,10(310) helices	No. of chain	Profile 3D Scores
L. donovani	7 / 19.2% 3(min)-18(max) Residues take part in formation of helices	2 /1.6% 3 residues	1	150.65
L. braziliensis	7/ 21.2 % 3(min)-17(max) Residues take part in formation of helices	4 / 3.9 % 3,5 residues	1	146.29
L. infantum	7/ 21.2 % 3(min)-17(max) Residues take part in formation of helices	4 / 3.9 % 3,5 residues	1	153.31
L. chagasi	8 /22.8 % 3 (min)-17(max) Residues take part in formation of helices	5/5.2 % 3-5 residues	1	154.79
L.major	8 /22.8 % 3 (min)-17(max) Residues take part in formation of helices	5/5.2 % 3-5 residues	1	154.79
L. mexicana	8 /22.8 % 3 (min)-17(max) Residues take part in formation of helices	5/5.2 % 3-5 residues	1	154.79

been detected for amino acid sequence of *L. braziliensis* and *L. mexicana*. In *L. braziliensis* few amino acids of catB protein participates in the formation of outer membrane. DNA repair as well as transportation activity are the novel function reported by us in different strains of catB protein of *Leishmania* with the help of support vector machine tool (http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi). Lipid binding property (88-92%) of cathepsin B has been predicted

Table 5. It shows the simulation of cathepsin B protein of six different *Leishmania* strains.

Cathepsin B Protein in six different strains of leishmania		CHARM M Energy Kcal/mol	Electrostatic Energy Kcal/mol	Initial energy Kcal/mol	Vander Wall Energy Kcal/mol
L. donovani	**Initial**	-19493.7	-22057		2007.61
	Final	-16436.7	-20840.7	-16472.9	-1732.12
L. braziliensis	**Initial**	-9705.3	-11186		-1288.92
	Final	-17110.1	-21271.7	-17137.8	-1768.44
L. infantum	**Initial**	-1580.56	-10111.5		5452.26
	Final	-16886.8	-21141.1	-16881.9	-1701.91
L. chagasi	**Initial**	14341	-10295.5		21547.8
	Final	-16816.9	-21211.9	-16825.1	-1796.56
L. major	**Initial**	14341	-10295.5		21547.8
	Final	-16944.8	-21190	-16960	-1795.06
L. mexicana	**Initial**	14341	-10295.5		21547.8
	Final	-16953.4	-21152.5	-16994.2	-1815.47

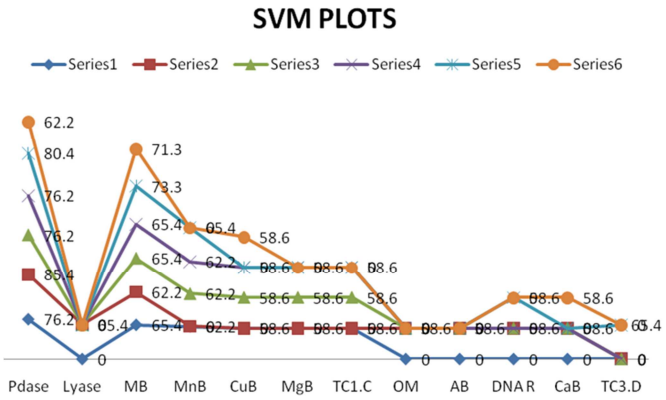


Figure 4. Comparative analysis of functional assignment of Cathepsin B protein in various strains by SVMProt. [Pdase→EC3.4, Hydrolases acting on peptide bonds, Lyase→ EC4.1 Lyase carbon-carbon lyases, MB→ Metal Binding, MnB→ Manganese binding, CuB→Copper Binding, MgB→ Magnesium Binding, Tc 1.c→ channels/pores-pore forming toxins (proteins & peptides), OM→ Outer membrane, AB→ Actin Binding, DNAR→ DNA Repair, CaB→ Calcium binding, Tc3-D→ Primary Active transporter oxidoreduction driven transporter.] (Series1→ L. donovani, Series2→ L.braziliensis, Series3→ L. chagasi, Series4→ L.major, Series5→ L. mexicana, Series6 → L. infantum).

in *L. donovani* (91.3%), *L. chagasi* (88.1 %), *L. major* (83.9%) and L. *infantum* (91.3%) is shown in Figure 4. From NCBI and EMBL, it is also known that Cathepsin B protein is a specific synthetic inhibitor protein which indicates that the inhibitor itself does not affect the growth of the parasites during the promastigote stages of the parasite.

3.4 *Predict Protein Server*

The amino acid sequence of cathepsin B protein of different *Leishmania* strains were submitted at protein predict server, to know whether there is presence of any post translation modification. In all six strains of *Leishmania* total eights post translational modification sites are observed. There are five aspargine glycosylation sites (N [^P] [ST] [^P]) from 18 amino acid to 159 amino acid in catB protein of six different strains of *Leishmania*. NFSV and NTTC glycosylation pattern is predicted at 28th and 159th residues in *L. chagasi, L. major, L. mexicana* where as in *L. donovani, L. braziliensis* (NFSV and NMST) glycosylation patterns were observed at 18th and 28th positions respectively. Glycosylation pattern (NSSK and NTTC) were observed at the 141th and 149th positions in catB protein of *L. donovani* and *L. mexicana*. Two protein kinase phosphorylation sites, one cAMP- and one cGMP-dependent with RRIS motif have been found at 88th position in *L. infantum*, *L. chagasi* and *L. mexicana* where as in *L. major* RRIS motif has been replaced by RRMS motif. In *L. donovani* and *L.braziliensis* at 78th position (motif RRIS), 260th position (RRGT motifs) are also present which imply that cAMP and cGMP dependent protein kinase phosphorylation occur at these sites. Five Protein kinase C activation sites have been found in three *Leishmania* strains i.e. *L. infantum*, *L. chagasi*

Table 6. Comparative analysis of different motifs of cathepsin B protein of six *Leishmania* strains. The motifs were predicted by Predicted Protein Server.

Motifs name	*L.donovani*	*L.braziliansis*	*L.infantum*	*L.chagasi*	*L.major*	*L.mexicana*
	PATTERN					
ASN_GLYCOSYLATION	18→NFSV 149→NTTC	18→NMST 159→NSTC	28→NMST 159→NSTC	28→NFSV 159→NTTC	28→NFSV 159→NTTC	28→NFSV 141→NSSK 159→NTTC
(N-glycosylation site) N[^P][ST][^P]						
CAMP_PHOSPHO_SITE	78→RRIS	260→RRGT	88→RRIS	88→RRIS	88→RRMS	88→RRIS
(cAMP- and cGMP-dependent protein kinase phosphorylation site) [RK]{2}.[ST]						
PKC_PHOSPHO_SITE	66→SDR 132→SDK 45→TPK 169→ SVK	3→SGK 25→SPR 45→SDK 76→SDR 179→SLR 208→SYK 278→SLK	3→SGK 76→SDR 142→ SDK 155→TPK 179→SVK	3→SGK 76→SDR 142→SDK 155→TPK 179→SVK	3→TGK 76→SDR 142→SEK 155→TPK 179→SVK	3→TGK 45→SEK 76→SDR 142→SSK 155→TPK 179→SIK
(Protein kinase C phosphorylation site) [ST].[RK]						
CK2_PHOSPHO_SITE	20→SVDE 43→TISE 141→TIYD 150→TTCE 155→SEMD	6 →SDEE 30→SAEE 40→TSFD 53→TISE 15 →TIYD 161→TCAD 217→TTGE	6 →SLEE 30→VDE 53→TISE 151→TIYD 160→TTC 165→SEMD	6 →SLEE 30→VDE 53 →TISE 151→TIYD 160→TTCE 165→SEMD	6 →SLGE 30→SVEE 53→TISE 151→TIYD 160→TTCE 165→SEMD	6→SLEE 30→SVEE 53→TIGE 160→TTCD
(Casein kinase II phosphorylation site) [ST].{2}[DE]						
TYR_PHOSPHO_SITE	154→KSEMDLVKY	171→KHKGEKSY	164→KSEMDLVKY	164→KSEMDLVKY	164→RSEMDLVKY	-
(Tyrosine kinase phosphorylation site) [RK].{2,3}[DE].{2,3}Y						
MYRISTYL	54→GSCWAI 95→GCYGGI 165→TSYSV 223→TQGGV 252 →SNECG 261→GVAGT	64→GSCWAI 105→CQGGI 233→GVQNGT 262 →GTDECG	64→SCWAI 105→CYGGI 175→GTSYSV 233→GTQGGV 262→SNECG 271→GVAGT	64→SCWAI 105 →CYGGI 175 →TSYSV 233→TQGGV 262 →SNECG 271→GGVAGT	64→GSCWAI 105→GCHGGGI 223→GTQDGV 262→GNNECK 271→GGVAGI	64→GSCWAI 105→GCYGGI 233→GVKDGI 262→GNDECG
(N-myristoylation site) G[^EDRKHPFYW].{2}[STAGCN][^P]						
THIOL_PROTEASE_CYS	50→QSNCGSCWAIAA	60→QSNCGSCWAIAA	60→QSNCGSCWAIAA	60→QSNCGSCWAIAA	60→QSNCGSCWAIAA	60→QSNCGSCWAIAA
(Eukaryotic thiol (cysteine) proteases cysteine active site) Q.{3}[GE].C[YW].{2}[STAGC][STAGCV]						
THIOL_PROTEASE_HIS	213→GGHAVKLVGWG	223→GGHAVKLVGWG	223→GGHAVKLVGWG	223→GGHAVKLVGWG	223→GGHAVKLVGWG	223→GGHAVKLVGWG
(Eukaryotic thiol (cysteine) proteases histidine active site) [LIVMGSTAN].H[GSACE][LIVM].[LIVMAT]{2}G.[GSADNH]						
PREDICTED SECONDARY STRUCTURE	H→13.33 E→23.33 L→63.33	H→15.71 E→18.21 L→66.07	H→15.36 E→21.79 L→62.86	H→15.36 E→22.50 L→62.14	H→16.07 E→19.29 L→64.64	H→17.14 E→20.00 L→62.86
GLOBULARITY	nexp=62 nfit =115 diff =47.00	nexp=172 nfit=119 diff=53.00	nexp=164 nfit=119 diff =45.00	nexp=164 nfit=119 diff=45.00	nexp=165 nfit=119 diff=46.00	nexp=170 nfit=119 diff=51.00
	nexp -number of predicted exposed residues					
	nfit -number of expected exposed residues					
	diff -difference nexp-nfit					

Table 7. It shows the prediction of disulphide bond and different motifs of cathepsin B protein of six different *Leishmania* strains.

Leishmania strains	Disulfide bond	Length	Sequence	Domains
L. donovani	10-188 111-218 123-166 126-162 158-208 159-326 196-222	178 107 43 36 50 167 26	AKSALCLVAVF – ITTEVCQPYPF EHWPMCVTISE –YDTPKCNTTCE RDQSNCGSCWA –ICGFGCYGGIP SNCGSCWAIAA – SCCFICGFGCY SNLLSCCFICG – DKYPPCPNTIY NLLSCCFICGF – RGSNECGIESG YPFGPCSHHGN KCNTTCEKSEM	1: 1-206 2: 207-340
L. braziliensis	3-111 123-166 126-326 140-218 158-222 159-188 196-208	108 43 200 78 64 29 12	XXXMRCYTKF-DKWPKCRTISE RDQSNCGSCWA-VCGMGCQGGIP SNCGSCWAIAA – RGTDECGIEST MSDRYCTVAGI – YDTPTCNSTCA GHLLSCCFVCG - TCNSTCADSHT HLLSCCFVCGM - LTSEVCQPYPF YPFPPCGHHTD - GKYPACPSTIY	1: 1-206 2: 207-340
L. infantum	10-188 111-218 123-166 126-162 158-208 159-326 196-222	178 107 43 36 50 167 26	AKSALCLVAVF – ITTEVCQPYPF EHWPMCVTISE - YDTPKCNTTCE RDQSNCGSCWA – ICGFGCYGGIP SNCGSCWAIAA – SCCFICGFGCY SNLLSCCFICG – DKYPPCPNTIY NLLSCCFICGF – RGSNECGIESG YPFGPCSHHGN - KCNTTCEKSEM	1: 1-206 2: 207-340
L. chagasi	10-188 111-326 123-166 126-162 158-208 159-208 196-222	178 215 43 36 50 49 26	AKSALCLVAVF – ITTEVCQPYPF EHWPMCVTISE – RGSNECGIESG RDQSNCGSCWA – ICGFGCYGGIP SNCGSCWAIAA – SCCFICGFGCY SNLLSCCFICG – DKYPPCPNTIY NLLSCCFICGF – YDTPKCNTTCE YPFGPCSHHGN - KCNTTCEKSEM	1: 1-206 2: 207-340
L. major	10-159 111-326 123-188 126-140 158-166 196-208 218-222	149 215 65 14 8 12 4	AKSALCLVAVF – NLLSCCFICGL EHWPMCLTISE – RGNNECKIESG RDQSNCGSCWA – IATEDCQPYPF SNCGSCWAIAA – ISDRYCTFGGV SNLLSCCFICG – ICGLGCHGGIP YPFDPCSHHGN - EKYPPCPSTIY YDTPKCNTTCE - KCNTTCERSEM	1: 1-206 2: 207-340
L. mexicana	10-162 111-218 123-166 126-326 158-208 159-188 196-222	152 107 43 200 50 29 26	TKSALCLVAVF – SCCFICGFGCY EKWPMCVTIGE – YNTPKCNTTCD RDQSNCGSCWA – ICGFGCYGGIP SNCGSCWAIAA – RGNDECGIESS TNLLSCCFICG - SKYPPCPNTIY NLLSCCFICGF – VTTELCQPYPF YPFGPCSHHGN - KCNTTCDNVEM	1: 1-206 2: 207-340

and *L. major* and their patterns are shown in (Table 6), but in *L. major* at first protein kinase C phosphorylation site serine has been replaced threonine. In *L. infantum*, *L. chagasi* and *L. major* six identical motifs of casein kinase II phosphorylation site were observed likewise at 6th and 30th positions codes SLGE and SVEE in L.major. In *L. donovani* and *L. braziliensis* and *L. mexicana* predicted 5, 7 and 4 different patterns are present respectively (Table 6). Three different tyrosine kinase phosphorylation sites (154→ KSEMDLVKY, 171 → KHKGEKSY and 164→RSEMDLVKY) have been observed in cathepsin B of *L. donovani, L. braziliensis* and *L. major* respectively. No tyrosine kinase phosphorylation site was detected in catB protein of *L. major.* Six N-myristoylation sites with same pattern have been observed in catB protein of *L. infantum, L. chagasi* and *L. major* but in *L. major* 175th pattern are absent. In *L. braziliensis* and *L. mexicana* having same four sites were observed. Amino acid composition of N-myristoylation site in *L. donovani* is completely different from all five strains of *Leishmania.* One motif of eukaryotic thiol (cysteine) proteases active site and eukaryotic thiol (histidine) proteases active site is present in different strains of *Leishmania* (Table 6).

Seven disulfide bonds formed between different amino acids. Two domains were identified in six *Leishmania* strains 1st domain is formed between 1-206 amino acids and 2nd domain is formed between 207-340 amino acids shown in Table 7.

4. Future Perspectives

Homology modeling of six different strains of *Leishmania* cathepsin B protein provided for the first time its 3-D structure model which could be tested for screening different molecules for the *Leishmania* specific cathepsin B inhibitory activity (docking analysis). The developed model showed good overall structural quality, and is validated using PROCHECK, WHATIF program. Prediction of different

functional sites like binding motifs, hydrolases sites, metal binding, glycosylation sites, protein kinase phosphorylation sites, N-myristoylation sites and different disulphide bridges are likely to be validated by experimental work. This knowledge could be used in biochemical studies to test the hypotheses of possible ligand binding sites. On the other hand, these experimental findings can then in turn be used to refine our models for virtual screening of chemical databases and rational drug design purposes. Advances in the field of insilico study will contribute to understanding between 3-D structure and ligand specificity of antileishmanial compound and it facilitate the development of various analogous of the presently available drug molecule on the basis of different binding sites of catB protein of different *Leishmania* strains.

Acknowledgements

This study was supported by a grant for setting up biomedical informatics centre from Indian Council of Medical Research (ICMR), Govt. of India. We acknowledge Dr. Sindhu Prava Rana for helping us in preparation of the manuscript. We thank Dr. Meera Singh, Scientist F of ICMR, New Delhi for helping us in setting up our biomedical informatics dept in RMRIMS, Patna, India.

References

1. P.J. Guerina, P. Olliarob, S. Sundard, M. Boelaerte, S.L. Croftf, P. Desjeuxg, M.K. Wasunnah, A. DM. Bryceson, *Lancet Infect.* Dis. 2 (2002) 494-501.
2. World Health Organization 1993 UNDP? World Bank/WHO 8, Leishmaniasis, Special Programme for Research and Training in Tropical Disease. Tropical Disease Research: Progress 1991-1992. Eleventh prgramme Report, 77-87.
3. B.L. Herwaldt, Lancet. 354 (1999) 1191-1199.
4. Y. Tselentis, A. Gilkas, B. Chaniotis, Lancet. 343 (1994), 1635.
5. R. Killick-Kendrick, 1979. Biology of *Leishmania* in phlebotomine sand flies. In Biology of the Kinetoplastida, W. Lumsden and D. Evans, editors. Academic Press, New York, USA.
6. J.M. Ribeiro, P.A. Rossignol, A. Spielman, Comp. *Biochem. Physiol.* 4 (1986) 683–686.
7. R. Charlab, J.G. Valenzuela, E.D Rowton, J.M. Ribeiro, *Proc. Natl. Acad. Sci.USA* 26 (1999) 15155–15160.
8. J. C. Mottram, D. R. Brooks, G. H. Coombs, Curr. *Opin. Microbiol.* 1 (1998) 455-460.
9. M. Knop, H.H. Schiffer, S. Rupp, D.H. Wolf, *Curr. Opin. Cell Biol.* 5 (1993) 990–996.
10. P.J. Berti, A.C. Storer, J. Mol. Biol. 246 (1995) 273-283.
11. J.S. Bond, P.E. Butler, Annu. Rev. Biochem. 56 (1987) 333-364.
12. K. Takio, T. Towatari, N. Katunuma, D.C. Teller, K. Titani , *Proc. Natl. Acad. Sci. USA* 80 (1983) 3666-3670.
13. H. Kirschke, A.J. Barrett, 1987. Chemistry of lysosomal proteases in Lysosomes: Their role in protein breakdown (Glaumann, H. & Ballard, F.J., eds), Academic *Press,* London. 193–238.
14. K.M. Karrer, S.L. Peiffer, M.E. DiTomas, *Proc. Natl.Acad. Sci. USA 90* (1983) 3063–3067.
15. M. Barral-Netto, A. Barrel, C. E. Brownell, Y. A. W. Skeiky, L. R. Ellingsworth, D. R. Twardzik, S. G. Reed, *Science.* 257 (1992) 545-548.
16. M. E. Wilson, B. M. Young, B. L. Davidson, K. A. Mente, S. E. McGowan, *J. Immunol.* 161 (1998) 6148-6155.
17. C. A. Hunter, H. Bermudez, H. Beernink, W. Waegell, J. S. Remington, Eur. J. Immunol. 25 (1995) 994-1000.
18. C. Bogdan, M. Rollinhoff, Parasitol. Today 15 (1999) 22-28.
19. C. Bogdan, J. Paik, C. Vodovotz, C. Nathan, *J. Biol. Chem.* 267 (1992) 23301-23308.
20. C. Bogdan, *Behring Inst. Res. Commun.* 99 (1997) 58-72.
21. J. Massague, Annu. Rev. Biochem. 67 (1998) 753-791.
22. J. S. Munger, J. G. Harpel, P. E. Gleizes, R. Mazzieri, I. Nunes, D. B. Rifkin, *Kidney Int.* 51 (1997) 1376-1382.
23. T. M. Chu, E. Kawinski, *Biochem. Biophys. Res. Commun.* 253 (1998) 118-134.
24. G.C. Sahoo, M. Rani, M.R. Dikhit, W.A Ansari, P.Das, Structural Modeling, Evolution and Ligand Interaction of KMP11 Protein of Different *Leishmania* Strains. J Comput Sci Syst Biol 2 (2009) 147-158.
25. J.D. Thompson, D.G. Higgins, T.J. Gibson, *Nucleic Acids Res.* 22 (1994) 4673-4680.
26. C.S.Ganesh, R.D.Manas, R.Mukta, D.Pradeep, Homology Modeling and Functional Analysis of LPG2 Protein of Leishmania Strains. J Proteomics Bioinform 0: (2009) 032-050.
27. A. D. MacKerell Jr., B. R. Brooks, C. L. III Brooks, L. Nilsson, B. Roux, Y. Won, M. Karplus, CHARMM: The Energy Function and Its Parameterization with an Overview of the Program, *The Encycl. Of Comp.Chem,* 1998, pp. 1271-1277.
28. Discovery Studio, Accelrys, San Diego, CA, USA.
29. C.Z. Cai, L.Y. Han, Z.L. Ji, X. Chen, Y.Z. Chen, *Nucleic Acids Res.* 31 (2003) 3692-3697.
30. G.C.Sahoo, M.R. Dikhit, P. Das, Functional assignment to JEV proteins using SVM. Bioinformation 3 (2008) 1-7.
31. P. Puntervoll, R. Linding, C. Gemünd, D.S. Chabanis, M. Mattingsdal *Nucleic Acids Res.* 31 (2003) 3625-3630.
32. B. Rost, G. Yachdav, J. Liu, *Nucleic Acids Research* 32 (2004) W321-W326.
33. A. Bairoch, P. Bucher, K. Hofmann, *Nucleic Acids Research* 25 (1997) 217-221.
34. A. Ceroni, P. Frasconi, A. Passerini, A. Vullo, *Bioinformatics* 20 (2004) 653-659.
35. D. Turk, M. Podobnik, R. Kuhelj, M. Dolinar, V. Turk, *FEBS Letters* 384 (1996) 211-214.
36. D. Musil, D. Zucic, D. Turk, R.A. Engh, I. Mayr, R. Huber, T. Popovic, V. Turk, T. Towatari, N. Katunuma, W. Bode, *EMBO J.* 10 (1991) 2321-2330.
37. M. Cygler, J. Sivaraman, P. Grochulski, R. Coulombe, A.C. Storer, J.S. Mort, *Structure* 4 (1996) 405-416.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.53

Assessing the Loss of Information through Application of the 'Two-hit Rule' in iTRAQ Datasets

Josephine Kilner[2], Liangjie Zhu[1], Saw Yen Ow[2], Caroline Evans[2,a], Bernard M. Corfe*[1,a].

[1]Department of Oncology, The University of Sheffield,The Medical School, Beech Hill Road, Sheffield, S10 2JF, UK; [2]Department of Chemical and Biological Engineering, ChELSI Institute, The University of Sheffield, Sheffield , Mappin Street, Sheffield, S1 3JD, UK; [a]These authors contributed equally to this work.

Received: 1 November 2010 **Accepted:** 17 November 2010 **Available Online**: 23 November 2010

ABSTRACT

High-throughput studies of complex protein mixtures using proteomic workflows typically employ tandem mass spectrometric analysis of peptides obtained by tryptic digestion. Protein identification is achieved by comparing the experimentally obtained peptide MS/MS spectra to theoretical spectra. Protein identifications based on peptide fragment sequences are often judged valid using the so called 'two-peptide' rule whereby any protein identified by sequencing of fragment ions must be justified by the identification of two sequence unique peptides from the same protein. This excludes proteins identified on the basis of a single peptide 'hit' (often termed a one-hit wonder, or OHW). Applying the 'two hit' stringency may result in the loss of potentially valuable meta-data: information yielded or consolidated by valid OHW proteins may be overlooked. This study tests the hypothesis that certain groups of OHW proteins (and thus related biological events or pathways) are more likely to be identified by single peptide due to various physical or biochemical characteristics (molecular weight and isoelectric point). We have undertaken analysis on data from three independent quantitative iTRAQ based proteomic studies of a human colon cell line and human colon tissue to correlate the differences between OHW and "valid" protein sets for molecular weight, isoelectric point and for associated biological pathways. The results show that there is a possible trend of inverse correlation between the pI value of a protein and the number of peptide hits for identification. Molecular weights range from 30-60 kDa. Pathway analysis using EBI-EMBL Reactome SkyPainter found that by excluding OHWs, several biological pathways were consistently not mapped, suggesting that exclusion of OHW potentially limits the understanding the biological processes potentially identified within the whole dataset. Future work should address strategies for evaluation of validity and reproducibility of these conclusions in other tissues.

Keywords: iTRAQ Mass-spectroscopy; One-hit Wonders; Protein Identification; GeneBio Phenyx search engine; Reactome-SkyPainter Pathway analysis.

Abbreviations

FDR false discovery rate; **iTRAQ LC MS/MS** Liquid-chromatography with tandem mass-spectroscopy using isobaric tag for relative and absolute quantitation; **MCP** Molecular & Cellular Proteomics; **MW** Molecular weight; **OHW** One-peptide wonders; **pI** Isoelectric point; **PSM** Peptide-spectrum matching; **SCFA** Short-chain fatty acidsReferences.

1. Introduction

The successful identification and relative quantification of the entire complement of proteins expressed by a whole cell or organism under certain conditions is a key goal of high throughput proteomics. The most common method of protein analysis is via the bottom-up approach, whereby enzymatically derived peptides (typically tryptic peptides) are analysed by mass spectrometry to determine their intact masses (MS) and the complement of ions from their dissociated fragments during gas phase fragmentation (MS/MS). Under these cases, detectable peptides (often called proteotryptic peptide) sequences from MS/MS spectra are matched against curated protein databases to give either positive or

***Corresponding author:** Dr. Bernard Corfe, Department of Oncology, The University of Sheffield,The Medical School, Beech Hill Road, SHEFFIELD, S10 2JF, UK; Email Adress: b.m.corfe@shef.ac.uk.

negative peptide identifications based on statistical and experimental supervised pattern matching criteria [1]. High throughput methods inevitably generate large datasets with spectra of varying quality. Peptide-identification tools and their underlying algorithms employ multivariate minimum score-threshold to segregate true/false results, with a 5% false discovery rate deemed as the upper limit by the Paris Guidelines (published in Molecular and Cellular Proteomics). A seemingly arbitrary 'two-peptide rule', was also originally recommended by 'MCP guidelines' [2], and is often applied under the collective assumption that more peptide identifications lead to a higher confidence protein identification. Whilst the two-hit rule remains one of the most applied unsupervised filters for high throughput proteomics techniques, there have been very few theoretical studies to support and describe the 'two-hit rule'. Interestingly, even the 'Molecular & Cellular Proteomics (MCP) guidelines' state that: "The two-peptide rule was discussed in the context of studies where there was little or no analysis done at all." [3,4]. The consequent flow-through from this filter implies that any protein identification that carries only single peptide evidence is viewed with a high level of uncertainty and is discarded from subsequent analysis [3,5].

In fact, protein identification that is supported by a single peptide with high confidence can theoretically be more valid than by two or more peptides with lower confidence. For example an analysis of human and Shewanella oneidensis datasets [4] showed that OHWs with a high peptide-spectrum matching (PSM) score were better for protein identification purposes than 'two-peptide proteins' with low PSM scores, especially when taking other proteomic information into consideration such as protein length. With the development of new technologies, the false positive rate (FPR) of protein identification is becoming lower. Global proteomic studies typically show FPRs below 5% for peptide identifications and more accurate methods for distinguishing false identifications are continuingly being found. These limitations apply, at least part, to the analysis of data derived from multiplex iTRAQ tandem mass-spectrometry. Briefly, iTRAQ is a gel-free approach that allows the identification and quantification of proteins across a diverse range of molecular weights (MW), pI values (isoelectric point), cellular locations and functional categories. Population of protein fragments (peptides) that are produced through bulk enzymatic digestion (typically tryptic) of the proteome are covalently labeled with a isotopically modified piperazine tag at both the N-terminus and amine side-chain amino acids [6]. As tags have been isotopically tampered to give isobaric masses, both non-intrusive relative and absolute quantifications for multiple comparisons can be made during the gas-phase dissociation (MS/MS) stage when mass different reporter tags are released for detection [7]. Subsequent database searches to match MS spectra with known peptide and protein sequences are then mostly analogous to other techniques, relying on search-engines and algorithms to filter identifications. As the analysis of a complex proteome is limited by the dynamic range of the sample and the practicing instrumentation, often every dataset analysis produces a significant number of OHWs. Apart from the limitation of low abundance and size of protein [8], certain groups of proteins may be consistently excluded by virtue of their physical characteristics and their dependency to crude extraction protocols. Thus, in cases where OHWs are the only detectable peptides of an enzymatic digest or are key regulators of a process, arbitrary removal of these potentially true positive identifications becomes particularly detrimental [4,9,10]; leading to a loss of potentially valuable information on their related pathways and mechanisms.

The outcome of an iTRAQ analysis is a list of differentially expressed proteins. A useful approach to aid interpretation of the resulting meta-information is metabolic pathway analysis. A pathway is defined as a set of interlinked, sequential biochemical steps that drives a cellular biological process and there are a number of packages (both proprietary and open-source) for bioinformatic analysis. Reactome SkyPainter [11] bases its analysis on a hypergeometric, or Fisher's exact test, a statistical significance test used when sample sizes are small which gives the probability of observing at least N genes from an event if the event is not overrepresented in the submitted list of genes. For further information see for example [12,13, 14,15].

This study forms part of our ongoing research into how short-chain fatty acids (SCFAs) act as chemo-preventives in colorectal cancer [16]. Three large datasets generated from multi-plex iTRAQ MS/MS experiments on colon cancer cells treated with SCFAs or from tissue from an in vivo study [16]. SCFAs are histone-deacetylase inhibitors (HDACi) which promote a range of transcriptional and post-translational changes in cells [16]. By using high-throughput proteomic techniques, including iTRAQ MS/MS, coupled with pathways analysis, a more global view of the range of actions of SCFA might be developed. The study also addresses the hypothesis that: "By discarding all 'OHW', valid protein identifications are missed with the potential loss of entire metabolic pathways. This has implications in the overall understanding of biological processes."

In order to assess the validity of this hypothesis statistical and pathway analyses were carried on the three datasets generated from iTRAQ quantification. The relationship between pI, Mw and peptide coverage with respect to peptide hits was described to characterize and potentially correlate relationship between 'valid' and OHW proteins. SkyPainter pathways analysis [17] was used to assess representation of pathways in the datasets, comparing 'valid' OHW depleted datasets with complete unfiltered (two-peptide rule) identifications.

2. Material and methods

2.1 Datasets and software

The three datasets used in our study were generated by 8-plex iTRAQ tandem mass spectrometry. Two datasets were based on HCT116 cell culture experiments, the third dataset

was derived from analysis of human colorectal biopsies [16]. The spectra were analysed using GeneBio's 'Phenyx' search engine. For peptide analysis and protein identification (http://www.genebio.com/products/phenyx) searches were undertaken against various publicly available datasets: NCBInr, UniProt and IPI. The results of these searches are available in Supplementary data 1a-c for each dataset. These contain details of all the proteins identified, including those by a single unique peptide, and related information on the number of validated unique peptide sequences on the protein, MS score, percent coverage, MW, pI value, etc. Table 1 gives a summary of the number of protein identifications and percentages of OHWs for each dataset. GraphPad Prism (GraphPad Prism 5 Demo) was used to analyze the correlation between MW and pI and the number of valid peptides for each protein identification and the means of each protein group.

shows non-significant negative linear trend. These data suggest that proteins with higher Mw are identified by more peptides, however taking the three analyses together, caution should be exercised in this interpretation and analysis of further datasets is required to establish such a relationship. The scatter-plots do show that the spread of MW is wider for OHWs than for proteins identified by more peptide peptides, consistent with observations by other groups [8] that typically >60% of proteins are identified by only one or two peptides.

Figure 1 panels d-f show the mean MW of proteins for the two groups: Dataset 1 (Fig. 1d) shows the mean MW of OHW's is significantly smaller than for proteins identified by more peptides (p-values <0.0001). However, no significant difference was observed for datasets 2 & 3 (Figs. 1e & f, p-values =0.2227 and 0.7612 respectively).

3.2 Correlation between pI value and number of valid peptides.

Table 1. *The protein numbers for the three datasets of our study.* The columns give the total number of protein identifications; the number identified by 1 or 2 valid peptides; the number identified by >2 valid peptides and the percentage of '1&2-hit' proteins.

Dataset	n, total number of proteins identified by Phenyx	Number of proteins identified by 1&2 peptides	Number of proteins identified by >2 peptides	Percent of proteins identified by 1&2 peptides
Dataset 1	262	163	99	62.2%
Dataset 2	599	419	180	69.9%
Dataset 3	209	188	21	90.0%

Microsoft Excel was used to construct surface plots for visualizing the relationship between all 3 parameters; MW, pI value, and valid peptide hits.

Pathway analysis was carried out using EBI-EMBL Reactome 'SkyPainter' (http://www.reactome.org/cgi-bin/skypainter2). This peer-reviewed and manually curated knowledgebase [14,17] includes biological pathway steps inferred to exist based on experimental data and provides an infrastructure for computation across the entire metabolic reaction network for multiple species, principally Homo sapiens. Pathways in Reactome are described as a number of molecular events that transform input physical entities into output entities in catalyzed or regulated pathways by other entities. By imputing a range of protein identifiers, 'SkyPainter' calculates which pathways are statistically over- or under-represented in a set of identifications, using a hypergeometric testing.

3. Results

3.1 Correlation between molecular weight and number of valid peptides.

We sought to establish whether or not there is a relationship, as previously suggested, between number of representative peptides and molecular weight of the protein [4]. Fig 1 shows the relationship between MW and the number of valid peptide peptides: Datasets 1 & 3 (Figs. 1a & 1c) show positive linear correlation, although this relationship is only significant (p-value <0.0001) for dataset 1. Dataset 2 (Fig. 1b)

Next we sought to investigate whether there is a relationship between pI and number of valid peptides. Figure 2a-c shows a consistent negative correlation between pI and the number of valid peptide peptides between the two groups for all 3 datasets. This correlation was significant (p-value <0.0001 & 0.0081 for datasets 1 & 2 respectively) but not for dataset 3 (Figs. 2a, b & c) which had a p-value of 0.0603. Although a consistent trend was observed, similar analysis of further datasets will be required to substantiate the significance of the correlation. The plots also show that proteins identified by fewer valid peptides show higher fluctuation of pI value, validating observations reported earlier [18].

3.3 Three-way interaction between Mw, pI and valid peptide number.

As there were trends to relationships between both Mw and peptides and pI and peptide number, the three-way interaction between these variables was explored using surface plots. These surface-plots allow the relationship between MW, pI value, and valid peptide hits to be visualized and qualitative interpretations to be drawn as well as provided data to support the interpretations. Figure 3a-c show surface plots of MW vs. pI by valid peptide hits. These were constructed by first grouping the results by valid peptide hits (OHW; 2 hits; 3 to 5 hits; etc.) and then calculating the mean MW for each group by pI band (pI 4 = ≥4.5 to <5.5; pI 5 = ≥5.5 to <6.5; etc.) Of note are the twin peaks in MW that occur at a similar position for each dataset; i.e. at an approximate pI of 6, with

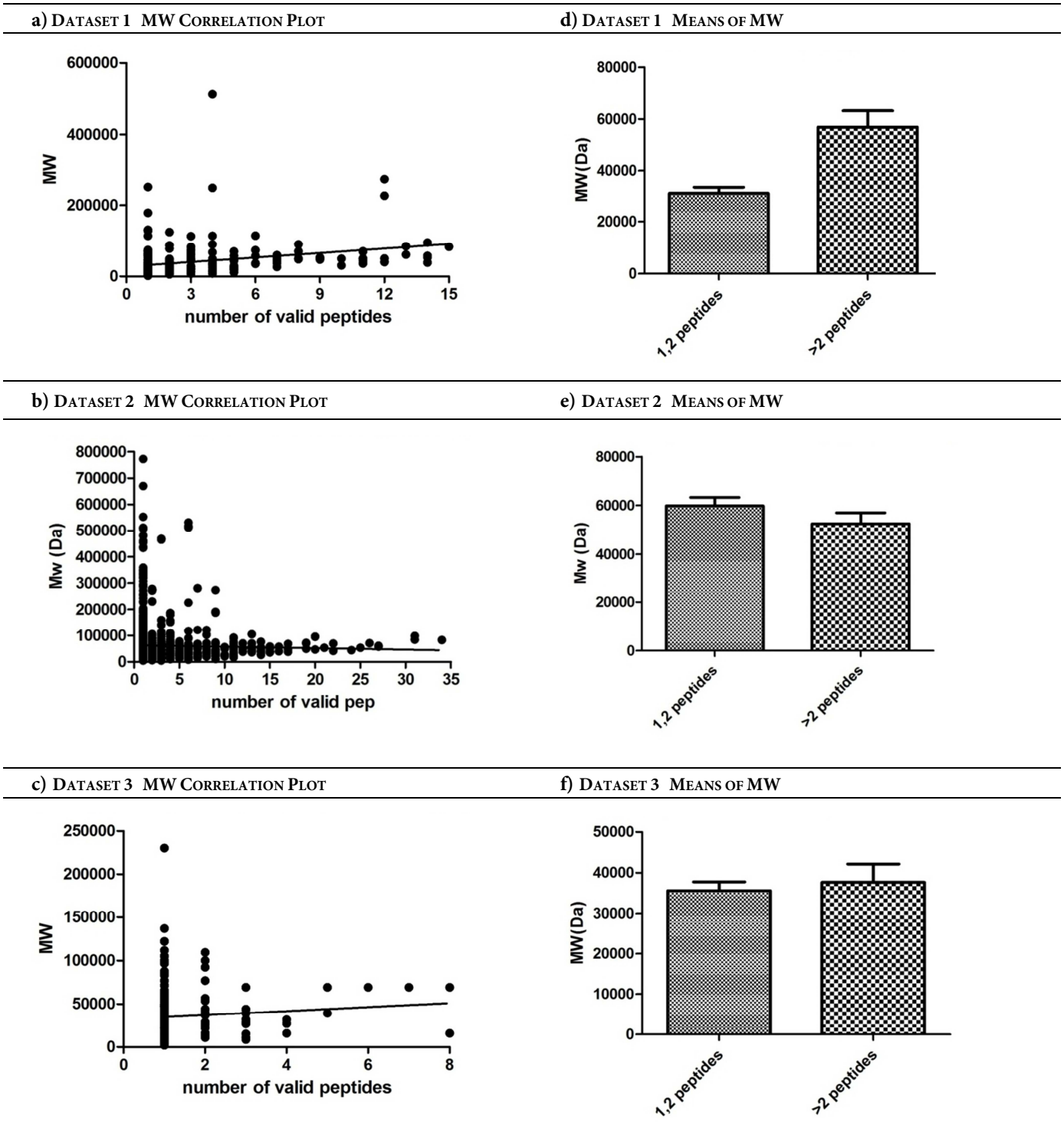


Figure 1. *The relationship between the M_W (Da) of a protein and the number of valid peptide hits in its identification.* Figs. 1a-1c are correlation plots for the three datasets respectively. With the exception of Dataset 1, there are no trends observed and no significant correlation. The results are as follows: **(a)** Dataset 1, slope= 4044±690 (n=262), p-value<0.0001; **(b)** Dataset 2, slope= -557.7±479.4 (n=599), p-value=0.2449; **(c)** Dataset 3, slope = 2319±1822 (n=209), p-value=0.2046. Figs. 1d-1f compares the mean M_W between the two protein groups: (i) proteins identified by 1or 2 peptide-hits; and (ii) proteins identified by >2 peptides. Again no relationship is observed and only Dataset 1 shows a significant difference (by unpaired t-test). The results are as follows: **(d)** Dataset 1, 1&2-hits (n=163) mean M_W=31,140±2,429 Da; >2-hits (n=99) mean M_W=56,830±6,305 Da, p<0.0001; **(e)** Dataset 2, 1&2-hits (n=419) mean M_W=59,769 ±3,470 Da; >2-hits (n=180) mean M_W=52,361±4,525 Da, p=0.2227; (f) Dataset 3, 1&2-hits (n=188) mean M_W=35,620±2,167 Da; >2-hits (n=21) mean M_W=35,640±4,512 Da, p=0.7612.

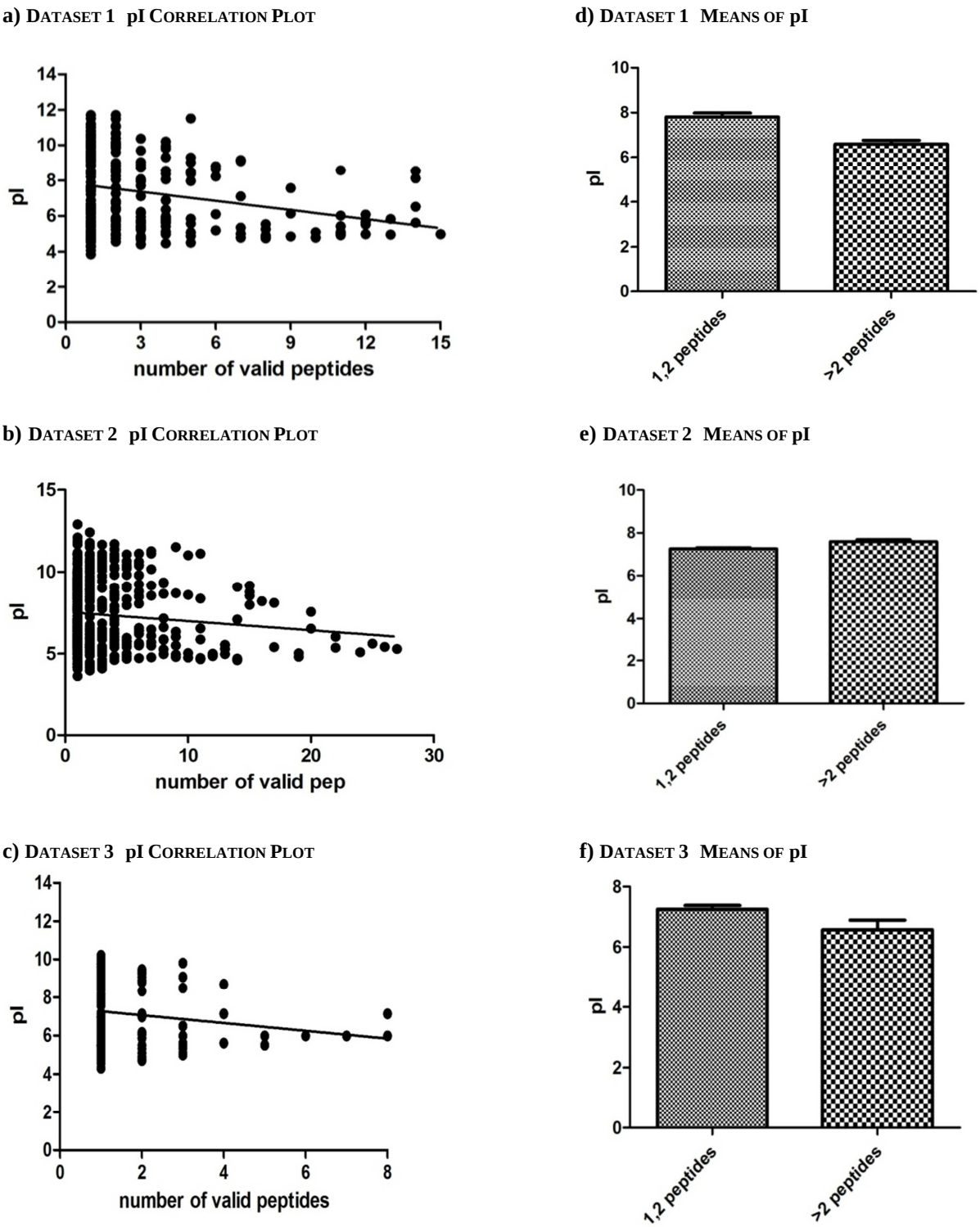
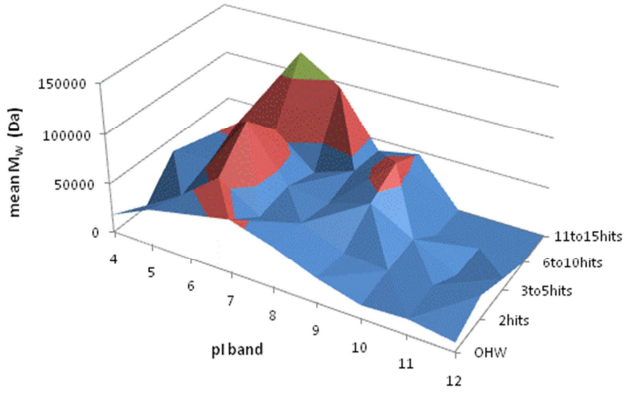
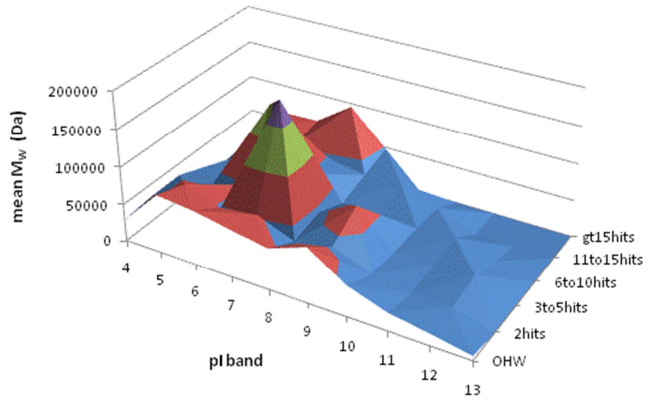


Figure 2. *The relationship between the isoelectric point (pI) of a protein and the number of valid peptide hits in its identification*. Figs. 2a-c are correlation plots for the three datasets respectively. Negative correlation is observed in all 3 datasets, although only Dataset 1 shows this to be significant. This suggests that peptides with high or low pI values are more likely to be represented by OHWs and so this group of proteins will potentially be excluded from any results. The results are as follows: (a) Dataset 1, slope= -0.1708±0.0411 (n=262), p-value<0.0001; (b) Dataset 2, slope= -0.5583±0.02101 (n=599), p-value=0.0081; (c) Dataset 3, slope = -0.2035±0.1078 (n=209), p-value=0.0603. Figs. 2d-f compare the mean pI values between the two protein groups: (i) proteins identified by 1or 2 peptide-hits; and (ii) proteins identified by >2 peptides. These results are inconsistent, only Dataset 1 shows a significant difference (by unpaired t-test) and there is no trend across the three datasets. The results are as follows: (d) Dataset 1, 1&2-hits (n=163) mean pI=7.803±0.1755 Da; >2-hits (n=99) mean pI=6.587±0.1758 Da, p<0.0001; (e) Dataset 2, 1&2-hits (n=419) mean pI=7.255±0.06110 Da; >2-hits (n=180) mean pI=7.584±0.1040 Da, p=0.0037; (f) Dataset 3, 1&2-hits (n=188) mean pI=7.248±0.1264 Da; >2-hits (n=21) mean pI=6.568±0.3215 Da, p=0.0854.

a) Dataset 1



b) Dataset 2



c) Dataset 3

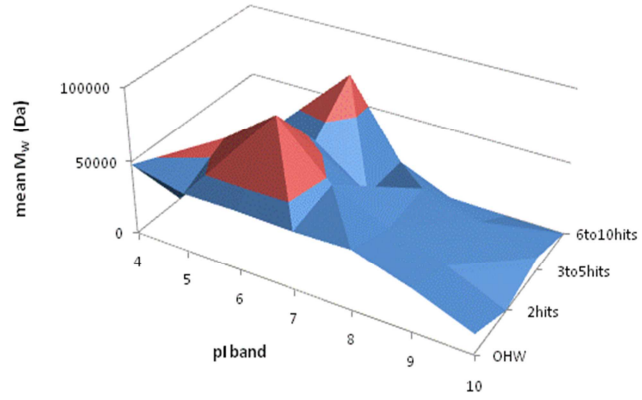


Figure 3. *Surface plots of M_W vs. pI by valid peptide hits for each of the 3 datasests.* The results have been grouped by valid peptide hits, then by pI band and the mean M_W for each group was calculated before the plots were constructed. The plots not only provide quantitative data but also provide a method of visualizing the relationships between these three parameters allowing qualitative interpretations and assessments to be made, for example the similarity between the position of the peaks in the plots for all 3 datasets and the observation that proteins in the OHW and 2-hit groups are only represented by low M_W proteins.

the first peak at 3 to 5 peptides and the second coinciding with the maximum peptide hits for each dataset (11-15 hits; >15 hits; 6-10 hits for datasets 1, 2 and 3 respectively).

3.4 Pathway Analysis.

EBI-EMBL's Reactome SkyPainter application [13] was used to compare events and pathways between the "OHW group" and "valid group" of proteins for each of the three datasets, where the "OHW group" includes proteins identified by 1 or 2 peptide hits and the "valid group" only contains proteins identified by more than 2 valid peptide hits. Each identified protein in the MS experiments was assigned to a reaction event and biochemical or metabolic pathway in the Reactome SkyPainter knowledgebase, (which contains a computationally-accessible human pathway network that has been manually curated as described in "Materials and Methods"). For user-submitted lists of protein identifiers, as carried out for the three datasets of this study, pathway over-representation analysis was performed and the results returned in the form of pathway trees colour-coded by probability. As such the relationships between our proteins and other complexes, reactions and pathways can be visualized. Data mining allows the pathway-trees to be expanded and the contributing reactions and events to be viewed. Hyperlinks allow the supporting literature to be accessed.

Full details of the SkyPainter analyses including event and pathway results for each dataset are provided in Supplementary information 2a-c. Supplementary-3 gives a summary of the event data with rows highlighted in red for results involving 2 or more of the datasets. Table 2 gives a list of events that were exclusively identified by the 1-2 peptide group for at least two of the datasets. Table 2 also gives the un-adjusted probability of seeing N or more genes in these events by chance (where a low probability indicates that the genes are statistically overrepresented in the pathway [15]). In total 7 events were identified that would have been lost in two or more of the datasets if the 1-2 peptide groups are discarded, with apoptosis being the only event consistently lost by all three datasets, as detailed in Table 2. Complete pathway-trees, including child-branches are given in Supplementary information 4a-c. The root-pathways are shown in Figure 4 and summarized in Table 3 as follows: (Table 3a) Pathway Trees unique to the group of proteins identified by only 1 or 2 peptides; (Table 3b) Pathway Trees unique to the group of proteins identified by more than 2 peptides; (Table 3c) Pathway Trees common to proteins in both groups.

A familiarization and working knowledge of the Reactome SkyPainter tool is important when interpreting results. For example, identification of an individual event that would be lost if OHW are discarded does not necessarily mean the pathway will also be lost because a pathway can be composed from many related events. We identified apoptosis as being a consistently lost event for all three datasets, but only a lost pathway for datasets 2 and 3 (see Table 3). Further investigation of the pathway-tree for dataset 1 (see Supplementary information 4a) shows this to be by virtue of related proteins

Table 2. *Events that were unique to proteins identified by only 1 or 2 valid peptides in two or more of the datasets.* These are events that would potentially be overlooked in any interpretation of the results if the 'two-hit rule is applied and all OHWs are excluded (full event lists can be found in the supplementary information).

Events unique to the group of proteins identified by only 1 or 2 peptide hits	Un-adjusted probability of seeing N or more genes in the event by chance			
	Dataset 1	Dataset 2	Dataset 3	Count
Apoptosis	0.00924	0.01356	0.02875	3
Cytosolic tRNA aminoacylation	0.04163	2.07E-05	-	2
Formation and Maturation of mRNA Transcript	0.00153	0.00060	-	2
mRNA Processing	0.00044	0.00015	-	2
Processing of Capped Intron-Containing Pre-mRNA	0.00016	0.00028	-	2
Processing of Capped Transcripts	0.00019	5.48E-05	-	2
Release of platelet cytosolic components	0.01105	-	0.00469	2

involved in the apoptotic execution phase (including Plectin; Importin subunit beta-1; and Lamin-A/C). Similarly, mRNA processing was identified as a lost event in OHWs in datasets 1 and 2 (see Table 2) but again was an identified pathway for dataset 1 (see Table 3) with further investigation revealing other related proteins and reactions including mRNA splicing (Supplementary information 4a).

4. Discussion

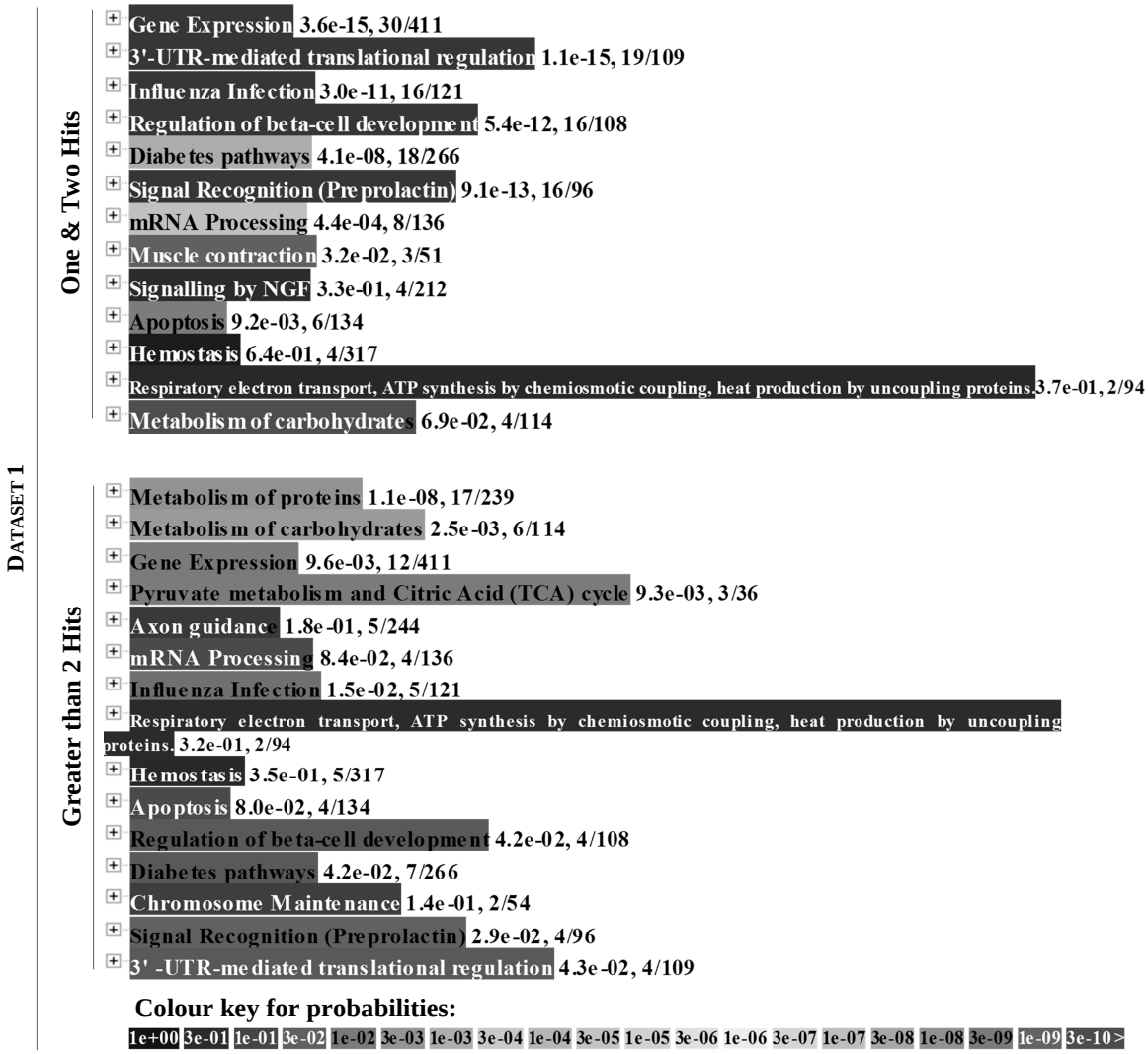
The ratio of MW to charge (m/z) is a key measurement metric in most MS protein identification and quantification. We had adopted the assumption, as proposed by others [7], that the number of trypsin cleavage sites increases with increased protein size, therefore the number of potentially detectable peptides should also increase accordingly. However our results (Fig. 1) showed an inconsistent correlation between MW and number of valid peptides and no significance and no consistency in the direction of difference between the mean MW of proteins identified by 1or 2 peptides or by more than 2 peptides. The lack of a strong and consistent correlation is surprising based on assumptions on purely stochastic grounds.

Using the second parameter, isoelectric point (pI), of the classes of proteins identified, we showed (Fig. 2) that although there was a trend, suggesting that proteins with high pI values produce fewer peptide fragments for identification, the results were only achieved significance for the first two datasets (Fig. 2a & b). The pI values of proteins identified in all three datasets are all between 4 and 14, with the exception of one protein having a pI<4. The results were unsurprising: a previous study [18] suggested that proteins with pI<4 have fewer arginine or lysine residues available for digestion by trypsin, thereby reducing the number of compatible peptides (for ion source protonation) available for positive ion mode (+ve MS) MS detection. Their study showed significant correlation in more than 95% of the proteins, with positive correlation for acidic proteins and inverse correlation for basic proteins. These data suggest the high fluctuations of pI values are therefore a consequence of protein length and amino acid composition, leading to broader and more sporadic pI shifts in shorter proteins.

Other studies [19] have also reported a higher proportion of negatively charged residues in peptides identified by MS with high confidence, with on average 16.8% of the residues in the high-scoring peptides being negatively charged, suggesting that the presence of acidic residues in a peptide may lead to more comprehensive and intense fragmentation of ions. EBI-EMBL Reactome SkyPainter was used for pathway and reaction event analysis (see Figure 4, Tables 2 & 3, and full details in the Supplementary information). A hypergeometric test is used to show events that are statistically overrepresented in the pathways. We had initially hypothesized that some pathways may be consistently lost if all OHWs are discarded. A pathway analysis of all three datasets has provided some evidence to support this and most importantly, a number of events were found to be uniquely identified by only one or two peptides in two or more of our datasets (see Table 2). While we have shown that one is able to validate and correlate extraneous layers of information given by traditionally discarded OHW proteins, we also recommend caution on the steps necessary to interpret these and other OHW data; for example, loss of event information does not necessarily mean loss of pathway information depending on which other proteins, events and reactions are involved, as discovered in our analysis of dataset 1 for the apoptosis and mRNA processing events, as discussed in detail earlier. We demonstrate that there are tangible benefits to compare and correlate OHW data, in turn minimizing overlooked pathways, and impairing the overall understanding of the biological process.

5. Concluding remarks

In this report, we have investigated the possibility of a relationship between MW and pI value of peptides and proteins


DATASET 1
One & Two Hits
Gene Expression 3.6e-15, 30/411
3'-UTR-mediated translational regulation 1.1e-15, 19/109
Influenza Infection 3.0e-11, 16/121
Regulation of beta-cell development 5.4e-12, 16/108
Diabetes pathways 4.1e-08, 18/266
Signal Recognition (Preprolactin) 9.1e-13, 16/96
mRNA Processing 4.4e-04, 8/136
Muscle contraction 3.2e-02, 3/51
Signalling by NGF 3.3e-01, 4/212
Apoptosis 9.2e-03, 6/134
Hemostasis 6.4e-01, 4/317
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, heat production by uncoupling proteins. 3.7e-01, 2/94
Metabolism of carbohydrates 6.9e-02, 4/114
Greater than 2 Hits
Metabolism of proteins 1.1e-08, 17/239
Metabolism of carbohydrates 2.5e-03, 6/114
Gene Expression 9.6e-03, 12/411
Pyruvate metabolism and Citric Acid (TCA) cycle 9.3e-03, 3/36
Axon guidance 1.8e-01, 5/244
mRNA Processing 8.4e-02, 4/136
Influenza Infection 1.5e-02, 5/121
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, heat production by uncoupling proteins. 3.2e-01, 2/94
Hemostasis 3.5e-01, 5/317
Apoptosis 8.0e-02, 4/134
Regulation of beta-cell development 4.2e-02, 4/108
Diabetes pathways 4.2e-02, 7/266
Chromosome Maintenance 1.4e-01, 2/54
Signal Recognition (Preprolactin) 2.9e-02, 4/96
3' -UTR-mediated translational regulation 4.3e-02, 4/109
Colour key for probabilities:
1e+00 3e-01 1e-01 3e-02 1e-02 3e-03 1e-03 3e-04 1e-04 3e-05 1e-05 3e-06 1e-06 3e-07 1e-07 3e-08 1e-08 3e-09 1e-09 3e-10 >

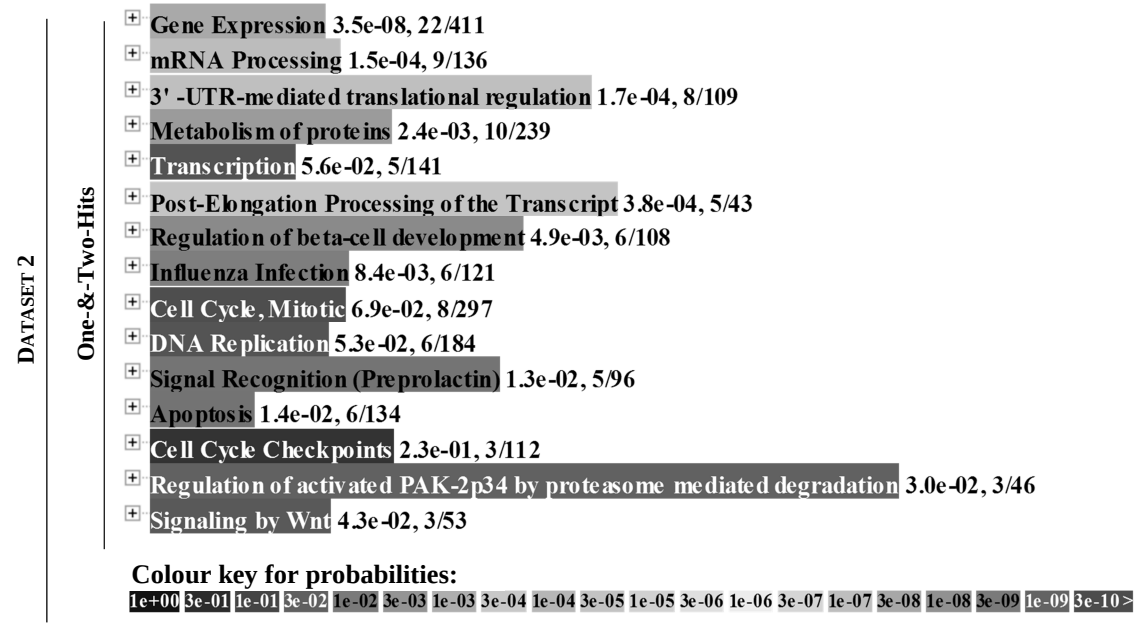

DATASET 2
One-&-Two-Hits
Gene Expression 3.5e-08, 22/411
mRNA Processing 1.5e-04, 9/136
3' -UTR-mediated translational regulation 1.7e-04, 8/109
Metabolism of proteins 2.4e-03, 10/239
Transcription 5.6e-02, 5/141
Post-Elongation Processing of the Transcript 3.8e-04, 5/43
Regulation of beta-cell development 4.9e-03, 6/108
Influenza Infection 8.4e-03, 6/121
Cell Cycle, Mitotic 6.9e-02, 8/297
DNA Replication 5.3e-02, 6/184
Signal Recognition (Preprolactin) 1.3e-02, 5/96
Apoptosis 1.4e-02, 6/134
Cell Cycle Checkpoints 2.3e-01, 3/112
Regulation of activated PAK-2p34 by proteasome mediated degradation 3.0e-02, 3/46
Signaling by Wnt 4.3e-02, 3/53
Colour key for probabilities:
1e+00 3e-01 1e-01 3e-02 1e-02 3e-03 1e-03 3e-04 1e-04 3e-05 1e-05 3e-06 1e-06 3e-07 1e-07 3e-08 1e-08 3e-09 1e-09 3e-10 >

Figure 4. *Root-level pathway trees for the three datasets***.** This shows the main pathways represented by the two protein groups: [(i) proteins identified by 1or2 peptide-hits; and (ii) proteins identified by >2 peptides]. (The complete pathway trees are given in the Supplementary information). Of note is how these pathways compare to the events shown in Table 2 indicating that caution needs to be taken when interpreting pathway and event data at face-value, as events not represented appear in the pathway tree by virtue of related events, e.g. the apoptosis event and the apoptotic execution phase which is a branch of the apoptosis pathway for the group >2-peptide hits in Dataset 1.

Table 3. Summary of the top-level pathways for the three datasets**.**

a) Pathways unique to proteins identified by 1 or 2 valid peptides

DATASET 1	DATASET 2	DATASET 3
Signalling by NGF	Apoptosis	Apoptosis
Muscle contraction	Cell Cycle Checkpoint	Axon guidance
	DNA Replication	Hemostasis
	mRNA Processing	Metabolism of proteins
	Post-Elongation Processing of the Transcript	Pyruvate metabolism and Citric Acid (TCA) cycle
	Regulation of activated PAK-2p34 by proteasome mediated degradation	
	Signalling by Wnt	
	Transcription	

b) Pathway Trees unique to the group of proteins identified by more than 2 peptides

DATASET 1	DATASET 2	DATASET 3
Axon guidance	Diabetes Pathway	
Chromosome Maintenance	Hemostasis	
Metabolism of proteins	Integration of energy metabolism	
Pyruvate metabolism and Citric Acid (TCA)	Interactions of the immunoglobulin super-	

cycle	family (IgSF) member proteins Metabolism of vitamins and cofactors	

c) Pathway Trees common to proteins in both groups

DATASET 1	DATASET 2	DATASET 3
3'-UTR mediated translational regulation	3'-UTR mediated translational regulation	Metabolism of carbohydrates
Apoptosis	Cell Cycle, Mitotic	
Diabetes Pathway	Gene Expression	
Gene Expression	Influenza Infection	
Hemostasis	Metabolism of proteins	
Influenza Infection	Regulation of beta-cell development	
Metabolism of carbohydrates	Signal Recognition (Preprotactin)	
mRNA Processing		
Regulation of beta-cell development		
Respiratory electron transport, ATP synthesis by chemioscopic coupling and heat production by uncoupling proteins		
Signal Recognition (Preprotactin)		

identified in iTRAQ experiments with the number of valid peptides to determine if any protein groups are consistently lost to any analyses, when OHWs are discarded. Although no relationship between MW was established, there was a trend towards negative correlation between pI value and number of peptide identification.

Pathway analyses highlighted several events that were only attributed to proteins identified by only one (OHW) or two peptides in two or more of our datasets.

While we acknowledge that the confirmation of our observations requires further analysis using orthogonal validation, for example by western blot analysis, we advocate the importance of the 'lost' information for a global interpretation of the data and therefore suggest that a more open approach should be taken when analyzing MS data since all candidate proteins/pathways will require verification. With the continued development of new technologies, software algorithms and bioinformatics tools, we believe the validation of OHW should become much more feasible [20, 21, 22].

6. Supplementary material

Supplementary data and information is available at:
http://www.jiomics.com/index.php/jio/rt/suppFiles/53/0

Supplementary 1a to c: Phenyx Protein Information for Datasets 1 to 3 respectively. Supplementary 2a to 2c: EBI-EMBL Reactome SkyPainter event and pathway results for Datasets 1 to 3 respectively. Supplementary 3: Summary of SkyPainter Events for the 3 Datasets. Supplementary 4a to c: Complete Pathway trees from SkyPainter for Datasets 1 to 3 respectively.

Acknowledgements

EPSRC funding, EPSRC grant GR/S84347/01, for the ChELSI initiative to support OSY and JK through a doctoral training grant to Chemical Engineering at the Life Science Interface at the University of Sheffield.

References

1. Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature 422, 198-207.
2. Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004). The need for guidelines in publication of peptide and protein identification data - Working group on publication guidelines for peptide and protein identification data. Molecular & Cellular Proteomics 3, 531-533.
3. Eddes, J.S., Kapp, E.A., Frecklington, D.F., Connolly, L.M., Layton, M.J., Moritz, R.L., and Simpson, R.J. (2002). CHOMPER: A bioinformatic tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies. Proteomics 2, 1097-1103.
4. Gupta, N., and Pevzner, P.A. (2009). False Discovery Rates of Protein Identifications: A Strike against the Two-Peptide Rule. Journal of proteome research 8, 4173-4181.
5. Bradshaw, R.A., Burlingame, A.L., Carr, S., and Aebersold, R. (2006). Reporting protein identification data - The next generation of guidelines. Molecular & Cellular Proteomics 5, 787-788.
6. Ernoult, E., Gamelin, E., and Guette, C. (2008). Improved proteome coverage by using iTRAQ labelling and peptide OFFGEL fractionation. Proteome Science 6.
7. Aggarwal, K., Choe, L.H., and Lee, K.H. (2006). Shotgun proteomics using the iTRAQ isobaric tags. Briefings in functional genomics & proteomics 5, 112-120.
8. Adkins, J.N., Varnum, S.M., Auberry, K.J., Moore, R.J., Angell, N.H., Smith, R.D., Springer, D.L., and Pounds, J.G. (2002). Toward a human blood serum proteome - Analysis by multidimensional separation coupled with mass spectrometry. Molecular & Cellular Proteomics 1, 947-955.
9. Pan, S.Q., Gu, S., Bradbury, E.M., and Chen, X. (2003). Single peptide-based protein identification in human proteome through MALDI-TOF MS coupled with amino acids coded mass tagging. Analytical Chemistry 75, 1316-1324.
10. Veenstra, T.D., Conrads, T.P., and Issaq, H.J. (2004). Com-

mentary: What to do with "one-hit wonders"? Electrophoresis 25, 1278-1279.

11. www.reactome.org (2009). Reactome: a human pathway database (Cambridge, UK, EMBL-EBI).
12. Deutsch, E.W., Lam, H., and Aebersold, R. (2008). Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. Physiological Genomics 33, 18-25.
13. Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., et al. (2005). Reactome: a knowledgebase of biological pathways. Nucleic Acids Research 33, D428-D432.
14. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., et al. (2009). Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Research 37, D619-D622.
15. Vastrik, I., D'Eustachio, P., Schmidt, E., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., et al. (2009). Reactome: a knowledge base of biologic pathways and processes (vol 8, pg 39, 2007). Genome Biology 10.
16. Drake, P.J., Griffiths, G.J., Shaw, L., Benson, R.P., and Corfe, B.M. (2009). Application of high-content analysis to the study of post-translational modifications of the cytoskeleton. Journal of proteome research 8, 28-34.
17. Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., et al. (2007). Reactome: a knowledge base of biologic pathways and processes. Genome Biology 8.
18. Kiraga, J., Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Biecek, P., Polak, N., Smolarczyk, K., Dudek, M.R., and Cebrat, S. (2007). The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. BMC Genomics 8, 163.
19. Benjamini, Y., and Hochberg, Y. (1995). Controlling The False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 57, 289-300.
20. Goodlett, D.R., Bruce, J.E., Anderson, G.A., Rist, B., Pasa-Tolic, L., Fiehn, O., Smith, R.D., and Aebersold, R. (2000). Protein identification with a single accurate mass of a cysteine-containing peptide and constrained database searching. Analytical Chemistry 72, 1112-1118.
21. Nesvizhskii, A.I., Keller, A., Kolker, E., and Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. Analytical Chemistry 75, 4646-4658.
22. Zhen, Y.J., Xu, N.F., Richardson, B., Becklin, R., Savage, J.R., Blake, K., and Peltier, J.M. (2004). Development of an LC-MALDI method for the analysis of protein complexes. Journal of the American Society for Mass Spectrometry 15, 803-822.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.52

Cell fractionation - an important tool for compartment proteomics

Maxie Rockstroh[1], Stephan A. Müller[1], Claudia Jende[1], Alexandra Kerzhner[1], Martin von Bergen[1,2], Janina M. Tomm*[1].

[1]Department of Proteomics, Helmholtz Centre for Environmental Research - UFZ, Permoser Str. 15, 04318 Leipzig, Germany; [2]Department of Metabolomics, Helmholtz Centre for Environmental Research - UFZ, Permoser Str. 15, 04318 Leipzig, Germany.

Received: 31 October 2010 **Accepted:** 22 November 2010 **Available Online**: 2 December 2010

ABSTRACT

In order to maximize coverage in proteome studies, a successful approach is the fractionation of cellular compartments. For providing evidence for the most reliable and efficient separation technique, we compared four different procedures for subcellular fractionation of Jurkat cells. The analysis of fractions by LTQ-Orbitrap yielded between 559 and 1195 unambiguously identified unique proteins. The assumed correct localization of the proteins was defined using Scaffold3 according to GO annotations, with the highest reliability (~80%) for the cytoplasmic fraction and the lowest (~20%) for the cytoskeletal fraction. This comparison revealed evidence for the efficiency of separating subcellular fractions and will thereby facilitate the decision on which procedure might be the best match to a specific research question and contribute to the emerging field of compartment proteomics.

Keywords: Subcellular compartments; Cellular fractionation; Protein localization; Mass spectrometry.

1. Introduction

In proteomics it is desired to obtain the largest possible coverage of the proteome of interest and especially to detect proteins of mediate or even minor abundance, too [1]. Beside the development of more and more sensitive mass spectrometers the most frequently applied approach for increased proteome coverage lies in the fractionation of the sample prior to analysis. This can be performed on the levels of subcellular compartments [2-4], proteins or peptides [5, 6] or a combination of different approaches [7]. The biologically most meaningful way is to separate subcellular compartments in order to preserve the linkage of proteins with the compartment in which they exert their activity. In many cases the biological relevance of a protein is closely linked to specific compartments and thereby it's influence on the whole phenotype of a cell.

Hence a great variety of methods for separating the subcellular compartments and subsequent proteome analysis have been developed (for review see [8]). Beside the coverage of the proteome, in praxis the hands-on time plays an important role for deciding in favor of a specific technique. Other criteria are reproducibility and in a few cases also high throughput capacity.

A well-established technique for separation of organelles is solely based on two different types of centrifugation, density velocity and density gradient centrifugation making use of differences in sedimentation coefficients and densities. With endpoint centrifugation, the membrane fraction of a broken cell can be obtained, regardless of the origin of the membrane [8]. Pellets resulting from a centrifugation scheme will stem mainly from the cytoplasmic membrane and only to lower percentages from organelles. A further sub-fraction that can be highly enriched by centrifugation contains the nuclei [9]. Due to their similarity in size but differences in density the remaining organelles like mitochondria, microsomes and lysosomes are often separated by density gradient centrifugation [10-12]. The centrifugation steps can be performed in buffers preserving protein structure and that are compatible with proteomic techniques like 2D-gel electrophoresis or LC-MS shotgun proteomics [8]. In summary, centrifugation schemes can be seen as recommended for enrichment of nuclei and membranes or for specific organelles like mitochondria, lysosomes and microsomes. Unfortunately, due to the

***Corresponding author:** Dr. Janina Tomm, UFZ Helmholtz Centre for Environmental Research, Department of Proteomics, Permoser Str. 15, 04318 Leipzig, Germany. Fax: +49-341-*2351787. Email* Address: Janina.Tomm@ufz.de.

nature of centrifugation, it is also time consuming and prevents high throughput.

In a more chemical orientated approach one can use a sequence of detergents with increasing solubilisation efficiency. Thereby a detergent like digitonin will be used to extract cytoplasmic proteins from a cell extract. The subsequent centrifugation will yield a highly enriched fraction of cytoplasmic proteins in the supernatant, whereas proteins from the pellet will be extracted by a stronger detergent like Triton X-100 [13]. There is a great variety in the sequence and choice of detergents described in other studies [14, 15]. Regrettably, this approach suffers from the wide variety of proteins and their interactions in turn leading to a modest specificity of extraction steps for subcellular compartments. Nevertheless, there are also some biologically highly relevant subcellular compartments like the proteome of the lipid rafts that can be extracted with high specificity [16].

In order to obtain high specificity and reproducibility while being cost- and time efficient, various combinations of physical and chemical methods using centrifugation and detergents have been developed. In addition, many protocols have been designed that lack ultracentrifugation and can be performed in volumes that are suitable for most widely distributed bench-top centrifuges, thereby increasing the high throughput capacity significantly.

Here we focused on the comparison of four different methods ranging from a rather simple separation into a soluble, mostly cytoplasmic fraction and an insoluble, mainly membranous fraction up to separation schemes leading to more than five different fractions. For three separations commercially available kits from Fermentas (ProteoJet Membrane extraction kit), Qiagen (Qproteome Cell Compartment Kit [17]) and Pierce (Subcellular Protein Fractionation Kit) were used. A fourth procedure was adapted from literature [18]. Hence we provide evidence for the decision on the most suitable separation for different purposes. It is noteworthy that the results might be cell line or tissue specific, so this has to be tested for the sample of choice. Here we focused on Jurkat cells, which serve as a cellular model for T helper-cells. They mimic important changes that also occur in native T-helper cells once they become stimulated. These processes lead to differential protein expression which has consequences in the cytoplasm, the nucleus and also in the membrane compartment.

With the development of shotgun mass spectrometry and data bases with predictions and reports on the subcellular distribution of proteins, a fast and reliable tool became available for testing the efficiency of the separation procedures. Again, in order to achieve optimal coverage and high reproducibility, a subfractionation was applied. The obtained fractions were applied to a SDS-gel and after a short run each lane was cut into three parts which were subjected to in-gel digestion. Measurement of the peptides by modern mass spectrometry revealed up to 670 proteins per fraction. For validating the results of subcellular fractionation approaches the number of several hundreds of proteins can be assumed to be sufficient to obtain a representative data set and for judging the success of the cellular fractionation.

In this study we provide evidence for the question which separation technique is the most favorable for a specific research question and approach. In addition to the achieved proteome coverage of subcellular compartments there are further requirements that need to be taken into account. For a specific research topic it might be helpful to use a combination of methods. The comparisons conducted here will help to facilitate proteomic research of subcellular compartments and organelles.

2. Material and methods

2.1 Cell culture

Jurkat T cells (clone E6-1, TIB-152, LGC Promochem, Wesel, Germany) were routinely maintained in RPMI-1640 medium (Biochrom AG., Berlin, Germany) containing 10% fetal bovine serum (Biochrom AG., Berlin, Germany), 1% L-Glutamine (Biochrom AG., Berlin, Germany), 1% streptomycin (100 mg/ml) / penicillin (100 U/ml) (PAA, Pasching, Austria) at an atmosphere of 5% CO_2, 95% humidity at 37 °C in a CO_2 incubator (MCO-18AIC, Sanyo Electric Co Ltd, Gunma-ken, Japan). Jurkat cells were cultured at 1 x 10^6 cells per ml medium. Cell viability and cell numbers were recorded by trypan blue exclusion.

2.2 Cell lysis and fractionation

All steps of the different fractionation methods were performed on ice using pre-chilled solutions unless noted otherwise. Centrifugation and incubation were carried out at 4 °C. If the composition of a buffer is not given, no further information was provided by the supplier. All fractions obtained were stored at -20 °C until further use. The fractionations were performed at least three times per method and the protein estimations were carried out in triplicates.

Method 1 (see also Fig. 1): Buffer 2 and 3 were supplemented with protease inhibitor solution (Roche, Mannheim, Germany) before use. Jurkat cells (5 x 10^6) were pelleted for 5 min at 250 x g and washed twice with 3 ml and 1.5 ml buffer 1, respectively. The cell pellet was resuspended in 1.5 ml buffer 2 by vortexing. The suspension was incubated for 10 min while continuously rocking. After 15 min centrifugation at 16,000 x g the supernatant 1 contained the cytosolic proteins. The pellet 1 was solved in 1 ml buffer 3 and the mixture was incubated for 30 min shaking at 1400 rpm in a thermomixer (Eppendorf, Hamburg, Germany). The suspension was centrifuged for 15 min at 16,000 x g. The supernatant 2 contained the membrane proteins, the cell debris containing pellet 2 was discarded. The protein determination for both fractions was carried out using the Bradford Quick Start Protein Assay according to the recommendations of the supplier (Bio-Rad Laboratories GmbH, München, Germany).

Method 2 (see also Fig. 1): All buffers were supplemented

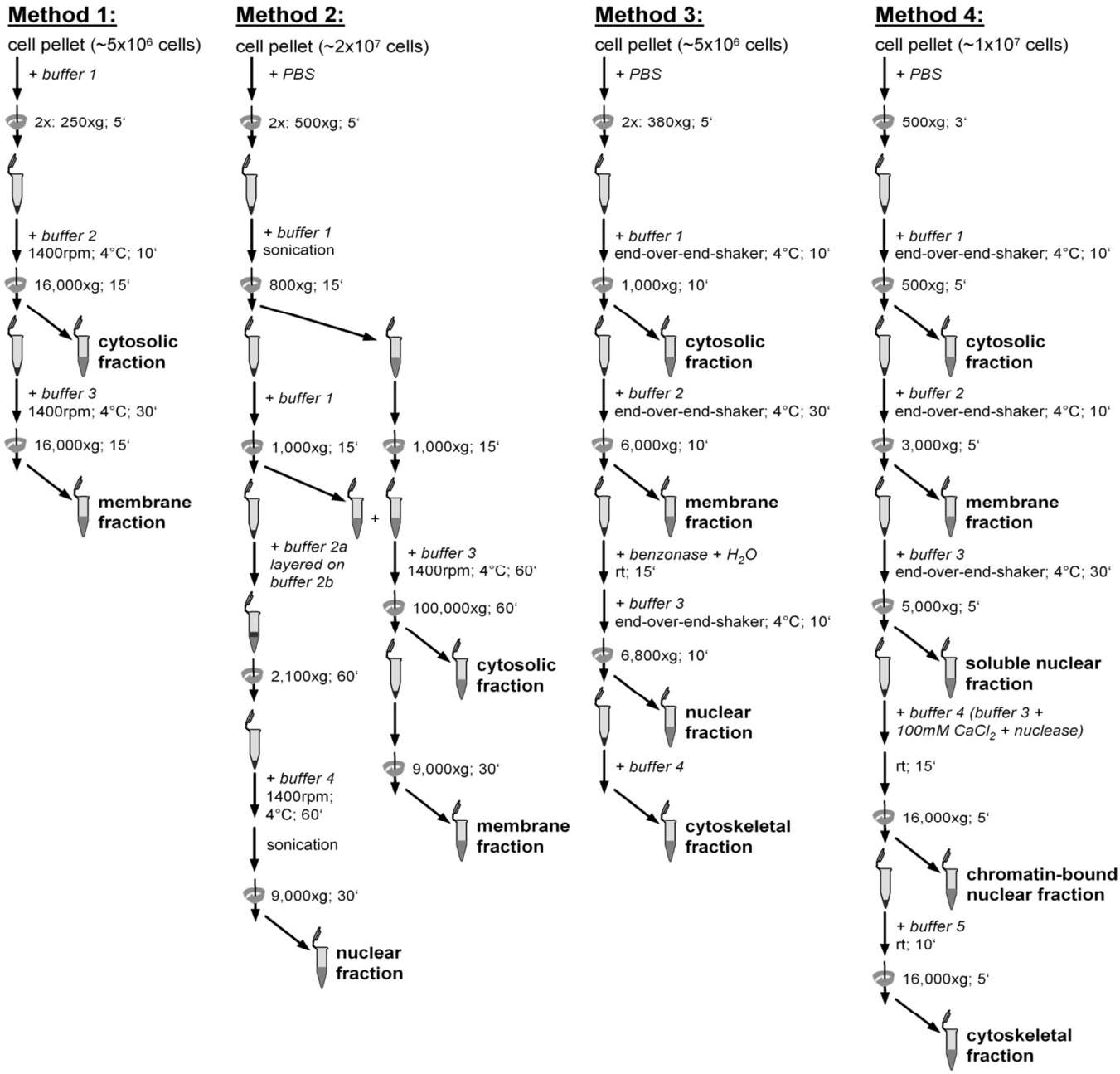


Figure 1. Schematic workflow. All centrifugation and incubation steps of the four different fractionation methods are shown (rpm is given for incubation in a thermomixer, x g for centrifugation).

with 1x protease inhibitor solution and 1 mM DTT directly before use. Jurkat cells (2 x 10^7) were washed twice with PBS and pelleted for 5 min at 300 x g. The cell pellet was resuspended in 1 ml buffer 1 (250 mM sucrose, 50 mM Tris-HCl, 5 mM $MgCl_2$) and cell lysis was performed by sonication on ice (3 times 10 s bursts with intensity ~40% and 30 s breaks). The suspension was centrifuged at 800 x g for 15 min and the pellet 1 was saved to isolate nuclei. The supernatant 1 was centrifuged again at 1,000 x g for 15 min. The obtained supernatant 2 was saved to isolate the cytosolic proteins, whereas pellet 2 was discarded.

The pellet 1 saved for isolation of the nuclei was dissolved in 1 ml buffer 1 and centrifuged at 1,000 x g for 15 min. The obtained supernatant 3 was added to the supernatant 2 for isolating cytosolic proteins and stored on ice until later. The pellet 3 was resuspended in 1ml buffer 2a (1 M sucrose, 50 mM Tris-HCl, 5 mM $MgCl_2$) and layered onto a 3 ml cushion of buffer 2b (2 M sucrose, 50 mM Tris-HCl, 5 mM $MgCl_2$). Afterwards centrifugation at 2,100 x g for 1 h was carried out. The pellet 4 was taken up in 500 µl buffer 4 (20 mM HEPES (pH 7.9), 1.5 mM $MgCl_2$, 0.5 M NaCl, 0.2 mM EDTA, 20% glycerol, 1% Triton X-100) and incubated 1 h

shaking at 1400 rpm and 4 °C in a thermomixer. Afterwards the suspension was sonicated again on ice (3 times 10 s bursts with intensity of ~40% and 30 s breaks) and centrifuged at 9,000 x g for 30 min. The supernatant 5 contained the nuclear proteins.

The pooled supernatants 2 and 3 were centrifuged for 1 h at 100,000 x g in an ultracentrifuge. The supernatant 6 contained the cytosolic proteins. The pellet 6 was solved in 0.5 ml buffer 3 (20 mM Tris-HCl, 0.4 M NaCl, 15% glycerol, 1.5% Triton X-100), incubated 1 h shaking at 1400 rpm and 4 °C and centrifuged at 9,000 x g for 30 min. The supernatant 7 contained the membrane proteins. The Lowry-DC-Protein Assay (Bio-Rad Laboratories GmbH) was used to determine the protein content of all fractions obtained with method 2.

Method 3 (see also Fig. 1): All buffers were supplemented with protease inhibitor solution before use. Jurkat cells (5×10^6) in a 1.5 ml reaction tube were pelleted for 5 min at 380 x g and washed twice with 1 ml PBS. The cell pellet was mixed with 1 ml buffer 1 and incubated for 10 min on an end-over-end shaker. The lysate was centrifuged at 1,000 x g for 10 min. The supernatant 1 contained the cytosolic proteins. The pellet 1 was resuspended in 1 ml buffer 2 and incubated for 30 min on an end-over-end shaker and centrifuged at 6,000 x g for 10 min. The newly gained supernatant 2 contained primarily membrane proteins. The pellet 2 was mixed with 20 µl distilled water containing 35% benzonase by gently flicking the bottom of the tube. After 15 min incubation at room temperature 0.5 ml buffer 3 was added and the suspension incubated for 10 min on an end-over-end shaker. The insoluble material was pelleted by centrifugation at 6,800 x g for 10 min. The supernatant 3 contained the nuclear proteins. The pellet 3 contained primarily cytoskeletal proteins and was resuspended in 250 µl room temperatured buffer 4. The protein content of all fractions was determined using the BCA Protein Assay Macro Kit (SERVA Electrophoresis GmbH, Heidelberg, Germany).

Method 4 (see also Fig. 1): All buffers were supplemented with protease inhibitor solution before use. Jurkat cells (1×10^7) were washed with PBS and pelleted for 3 min at 500 x g in 1.5 ml reaction tubes. The cell pellet was solved in 1 ml buffer 1 and incubated for 10 min on an end-over-end shaker. The lysate was centrifuged at 500 x g for 5 min. The supernatant 1 contained the cytosolic proteins. The pellet 1 was mixed with 1 ml buffer 2, vortexed and incubated for 10 min on an end-over-end shaker. After centrifugation at 3,000 x g for 5 min, the obtained supernatant 2 contained primarily membrane proteins. The pellet 2 was dissolved in 0.5 ml buffer 3, vortexed and incubated for 30 min on an end-over-end shaker. Following centrifugation at 5,000 x g for 5 min the supernatant 3 contained soluble nuclear proteins. Buffer 4 was used at room temperature and prepared by adding 25 µl of 100 mM $CaCl_2$ and 15 µl of micrococcal nuclease to 0.5 ml buffer 3. 0.5 ml buffer 4 was added to the cell pellet 3, vortexed and incubated for 15 min at room temperature. The mixture was vortexed 15 s and centrifuged at 16,000 x g for 5 min. The supernatant 4 contained chromatin-bound nuclear proteins. The pellet 4 was resuspended with 0.5 ml buffer 5, vortexed and incubated for 10 min at room temperature. After centrifugation at 16,000 x g for 5 min the supernatant 5 contained the cytoskeletal proteins. The protein content of all fractions was determined using the BCA Protein Assay Macro Kit following the manufacturer's instructions (SERVA Electrophoresis GmbH, Heidelberg, Germany).

2.3 1D-gel electrophoresis

20 µg protein of each fraction were precipitated 15 min at -20 °C by addition of a 5-fold volume of ice cold acetone. The precipitates were centrifuged at 16,000 x g and 4 °C for 10 min and the supernatant was discarded. The dried pellets were dissolved in SDS-sample-buffer (62.5 mM Tris-HCl (pH 6,8), 10% glycerol, 2% SDS, 5% mercaptoethanol, 0.05% bromophenol blue) and separated by SDS-PAGE on a 4% stacking gel and 12% separation gel run according to standard laboratory procedures. For visual control of successful separation the gels were stained with Coomassie Brilliant Blue G250 after electrophoresis. For protein analysis and MS identification the proteins were allowed to enter only for about 2-3 cm into the gel and cut into 3 gel slices per sample after short staining with Coomassie solution.

2.4 Trypsin digestion and analysis by LC-MS/MS

The gel slices were destained with 50% methanol containing 5% acetic acid. After reduction with 10 mM DTT, proteins were alkylated with 100 mM iodoacetamide and then digested overnight at 37 °C using sequencing grade trypsin (Roche Applied Science, Mannheim, Germany). All membrane fraction containing gel slices were digested in a trypsin solution containing 30% methanol (except method 4). The resulting peptides were extracted two times from the gel with 5% formic acid and 50% acetonitrile. The combined extracts were evaporated, the residual peptides were dissolved in 0.1% FA and the solution was desalted by using C18-StageTips (ZipTipC18, Millipore Corporation, Billerica, MA, USA).

A nano-HPLC system (nanoAquity, Waters, Milford, MA, USA) coupled to a an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) via a nano electrospray ion source (TriVersa NanoMate, Advion, Ithaca, NY, USA) was used for LC/MS/MS analysis. Chromatography was performed with 0.1% formic acid in solvents A (100% water) and B (100% acetonitrile). Samples were injected on a trapping column (nanoAquity UPLC column, C18, 180 µm×20 mm, 5 µm, Waters) and washed with 2% acetonitrile containing 0.1% formic acid and a flow rate of 15 µl/min for 8 min. Peptides were separated on a C18 UPLC column (nanoAcquity UPLC column, C18, 75 µm×100 mm, 1.7 µm, Waters). Peptide elution was conducted using a gradient from 2 - 70% solvent B (0 min - 2%; 5 min - 6%; 45 min - 20%; 70 min - 30%; 75 min - 40%; 80 min - 70%) with a flow rate of 300 nl/min.

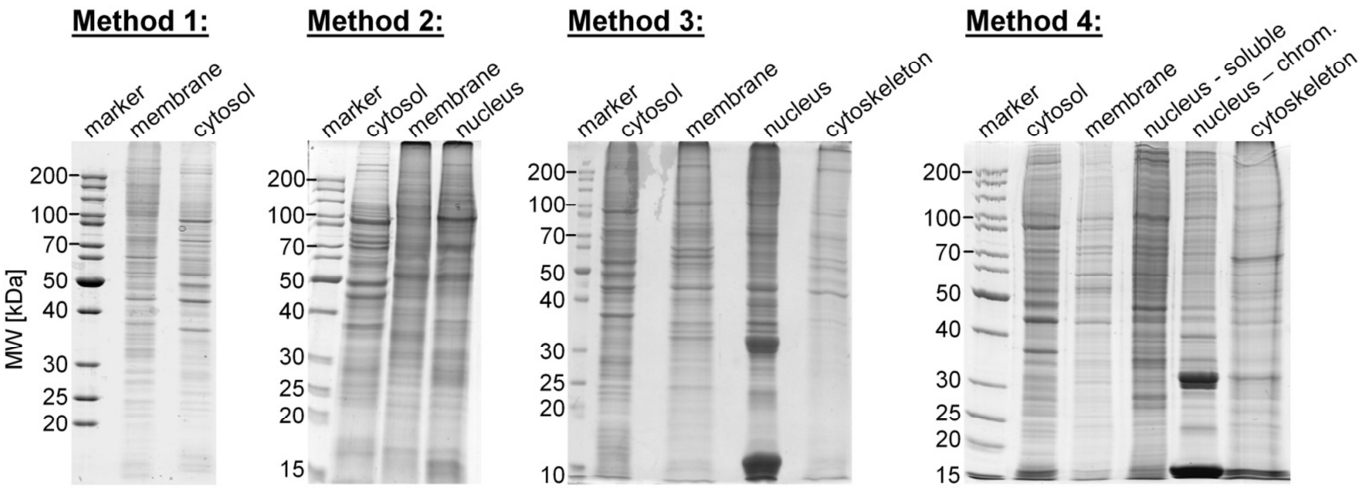


Figure 2. 1D-gels showing the different subcellular fractions. For initial evaluation of the fractions obtained by the four different methods, 20 µg of each protein fraction were separated in a 12% SDS-Gel and stained with colloidal Coomassie. The marker is located on the left hand side of each gel (nucleus-chrom. = chromatin-bound nuclear fraction).

Full scan MS spectra (from 400-1500 m/z, R = 60000) were acquired in positive ion mode in the LTQ-Orbitrap.

Peptide ions exceeding an intensity of 3000 were chosen for collision induced dissociation within the linear ion trap (isolation width 4 m/z, normalized collision energy35, activation time 30 ms, activation q = 0.25). For MS/MS acquisition, a dynamic precursor exclusion of 2 min was applied.

2.5 Data analysis of the mass spectrometric results

MS/MS samples were analyzed by Proteome Discoverer (version 1.0; Thermo Fisher Scientific, San Jose, CA, USA) using the MASCOT search algorithm (version 2.2.06; Matrix Science, London, UK) [19]. Mascot was set up to search a reverse concatenated database of all human proteins annotated in the SwissProt database (version 10/07/2010) assuming the digestion enzyme trypsin. Mascot was searched with a fragment ion mass tolerance of 0.5 Da and a parent ion tolerance of 5 ppm. Carbamidomethylation of cysteine was specified as a fixed modification. Oxidation of methionine and acetylation of the protein n-terminus were specified as variable modifications.

Scaffold 3 (version Scaffold 3_00_03, Proteome Software Inc., Portland, OR, USA) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they exceeded specific database search engine thresholds. Mascot identifications required at least ion minus identity scores of greater than -5 and ion scores of greater than 15. Protein identifications were accepted if they contained at least 2 identified peptides. Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. False discovery rate of proteins was determined to be lower than 0.2% for all samples. Gene ontology annotations were obtained from the EBI GO database (www.ebi.ac.uk/GOA/, version 10/08/2010).

3. Results and Discussion

3.1 Fractionation of Jurkat cells

The workflow of the four different methods used to fractionate Jurkat cells into several cellular compartments is shown schematically in Fig. 1. In method 1, 3 and 4 commercially available kits were used, whereas method 2 uses an adapted protocol from Nature Protocols [18]. All methods rely on cell lysis through sequential addition of different buffers to the cell pellets followed by incubation and centrifugation at different speeds. In method 2 sonication is additionally used to lyse the cells. From method 1 only two different fractions, cytosol and membrane, were obtained. In addition to the three fractions prepared with method 2 – cytosol, membrane and nucleus, a fourth cytoskeletal fraction can be separated with method 3. With method 4 even five different subcellular fractions can be isolated: cytosol, membrane, cytoskeleton, with the nuclear fraction further split into soluble and chromatin-bound nuclear fraction. Method 1 is least time consuming, with about 1.5 hours needed for the fractionation. In approximately 2 hours a fractionation with method 3 or 4 is completed. With at least 3.5 hours of work method 2 is the longest protocol of all four. In addition, method 2 is the most complicated protocol because there are two lines of work steps which have to be performed in parallel while all other methods require only one straight workflow. Moreover, an ultracentrifuge with acceleration up to 100,000 x g is needed for method 2, while a normal table-top centrifuge with up to 16,000 x g is sufficient for all other methods used. Nevertheless, all buffers for method 2 can be prepared in the lab and no expensive kit is needed and the largest number of protein identifications was obtained.

The total amount of obtained protein differed for the various methods (Tab. 1) from 0.78 mg to 3 mg per 1 x 10^7 cells, ranging between 0.5 and 1.57 mg for the cytoplasmic fraction and 0.08 to 0.92 mg for the nuclear fraction. This shows that

Table 1. Protein amounts obtained per 1 x 107 cells in each fraction.

Amount of protein obtained per 10^7 cells [mg]

Method / Fraction	1	2	3	4
Cytosol	0.705	0.495	0.496	1.571
Membrane	0.630	0.208	0.135	0.362
Nucleus	–	0.079	0.183	–
Nucleus - soluble	–	–	–	0.520
Nucleus - chromatin-bound	–	–	–	0.401
Cytoskeleton	–	–	0.044	0.148
Total amount of protein	1.335	0.782	0.858	3.002

Table 2. Number of proteins identified in the subcellular fractions.

Method / Fraction	1	2	3	4	Ø proteins identified / method
Cytosol	414	657	599	620	573
Membrane	249	458	352	523	396
Nucleus	-	603	258	-	431
Nucleus - soluble	-	-	-	670	670
Nucleus - chromatin-bound	-	-	-	370	370
Cytoskeleton	-	-	618	64	341
Total number of identified proteins	559	1231	1126	1195	1028

there is a rather wide variance in efficiency of the protein isolation. This should also to be taken into account when choosing the fractionation method combinable with the protein detection method used afterwards.

3.2 1D-gel electrophoresis

A first overview of the successful protein separation by the different fractionation methods was obtained by SDS-PAGE. All fractions gained using one method show clearly different band patterns, whereas the same subcellular fractions from different methods have some resemblance in their protein patterns (Fig. 2).

All cytosolic fractions show a comparable band pattern (e.g. five strong bands, of which one is at ~90 kDa, one slightly above 50 kDa, two between 40 and 50 kDa and one at ~38 kDa). Likewise the membrane fractions of method 1, 3 and 4 have a similar band pattern showing a more distinct band at approximately 60 kDa, whereas the separated membrane proteins of method 2 seem to run at slightly different heights. The nuclear fraction from method 2 has as well only partial similarities to the nuclear fractions of methods 3 and 4. The nuclear fraction from method 3 and the nuclear chromatin-bound fraction from method 4 show both two very prominent bands at ~15 and ~30 kDa. These bands are likely to represent histones. The soluble nuclear fraction from method 4 shares a stronger band at ~45 kDa with the nuclear fraction from method 3. As this band is also present in the chromatin-bound fraction, this protein might either be only loosely bound to the chromatin, or, more likely, is not completely separated from the chromatin-bound fraction.

3.3 Identification of proteins

The MS/MS data were analyzed by Proteome Discoverer using the MASCOT search algorithm. The MS/MS based peptide and protein identifications were validated by Scaffold 3. For evaluation of method 4 the two nuclear fractions were combined.

In the cytosolic fractions an average of 573 proteins was identified by all methods (Tab. 2). In the membrane fraction the amount of identified proteins varies a lot between the different methods. With method 1 only 249 proteins were found, whereas 523 proteins were identified with method 4. With method 2 more than the double amount of proteins (603) could be identified in the nucleus compared to method 3 (258). The two different nuclear fractions, soluble and chromatin-bound, obtained with method 4 yielded in 670 and 370 identified proteins, respectively, leading to 750 identified proteins for the nucleus in total (Fig. 3). The amount of cytoskeletal proteins identified with method 3 and 4 ranges from 64 proteins identified with method 4 and up to 618 with method 3. The total numbers of identified proteins were in the same range (between 1126 and 1231) for method 2, 3 and 4 while for method 1 only 559 proteins could be identified in total. Altogether, only the amount of identified proteins in the cytoplasmic and the membrane fractions are comparable within all methods. All methods differ significantly in the amount of proteins identified per fraction as well as in the amount of protein isolated in total.

3.4 Enrichment factor of different fractionation methods

To get a deeper insight into how efficiently each fractionation method worked out, the overlap and intersections in cytosolic, membrane and nuclear fraction were determined and plotted in venn diagrams (Fig. 3). For this aim the two nuclear fractions of method 4, soluble and chromatin-bound, were combined. The most proteins identified in two overlapping fractions were found in cytosol and membrane for method 1 and 3, whereas method 2 and 4 show the biggest overlap in the membrane and nuclear fraction.

Disregarding method 1, because it only yielded two frac-

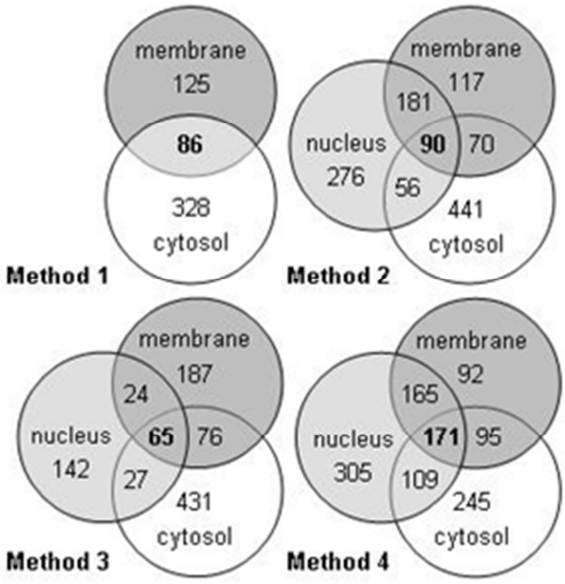


Figure 3. Overlap of proteins identified in the different subcellular fractions. For each of the fractionation methods used, a venn diagram was generated showing the overlap of the proteins identified in more than one fraction.

tions, the most proteins identified in only one fraction could be found with method 3 (80%). 68% of the identified proteins were found in only one fraction with method 2. Method 4 showed the smallest part of proteins identified in only one fraction (54%), while 46% of the identified proteins in this method were found in two or three of the fractions.

In this experiment Gene Ontology (GO) annotations were used by the evaluation program Scaffold 3 to analyze the subcellular localization of each protein identified in the samples. If the proteins identified in one fraction were supposed to be in that fraction according to the GO annotations, they were counted as proteins isolated in the 'correct' fraction. To compare how efficient each of the four fractionation methods fractionated the cells, the number of properly isolated proteins in each fraction was calculated. The percentage of the correctly separated proteins out of the total number of identified proteins in each fraction was calculated, too (Fig. 4). The cytosolic fraction was among all four methods the fraction with the most accurately isolated proteins (between 357 and 657 proteins) and comparable percentages about 80%. Between ~30 and 42% of the proteins found in the different membrane fractions where isolated correctly, leading to 74 till 188 isolated proteins in the 'correct' fraction in total. For the nuclear fraction 230 up to 345 nuclear proteins could be identified. The percentage of correct nuclear proteins from method 3 was very high with 90%, whereas method 4 showed a high amount of properly isolated proteins because of its two different nuclear fractions. Taking a closer look at transcription factors, there were 12 different ones detected using method 1 and 27 to 32 using method 2 to 4. With method 3 more appropriately isolated cytoskeletal proteins could be identified than with method 4, but the percentage is very low for both methods. The high false positive rate is likely due to the solubilisation of most of the proteins of the last cell pellet, where surely proteins of not completely dissolved membranes or other cellular compartments were inside.

3.5 Discrepancies between the predictions of the evaluation program and the measurements

The Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is a highly abundant protein, which accounts for 10 to 20% of the total cellular protein. It is commonly known as a glycolytic enzyme located in the cytoplasm with a key role in energy production [20]. By intensive research it became obvious that the GAPDH is in reality a multifunctional protein with diverse subcellular localizations in mammalian cells. The GAPDH can be found in the membrane, where it promotes endocytosis and membrane fusion and therefore vesicular secretory transport [21, 22]. Furthermore GAPDH is involved in the nuclear transport of RNA [23] and has the ability to activate the transcription in neurons [24]. Other functions in the nucleus are the assistance in DNA replication and DNA repair [25]. Due to the modulation of the cytoskeleton GAPDH can also be found in the cytoskeletal fraction [26, 27]. Thus the GAPDH can have not only a cytosolic, but also a membrane, nuclear and/or cytoskeletal localization.

According to the GO annotations the GAPDH is located only in the cytoplasm and membrane. This is contradictory to the various localizations described by the literature. In this experiment the GAPDH was found in all fractions obtained with method 2 and 4. With method 3 the enzyme was identified in the cytoplasmic, membrane and cytoskeletal fraction. For all of these three methods the localization in nucleus and

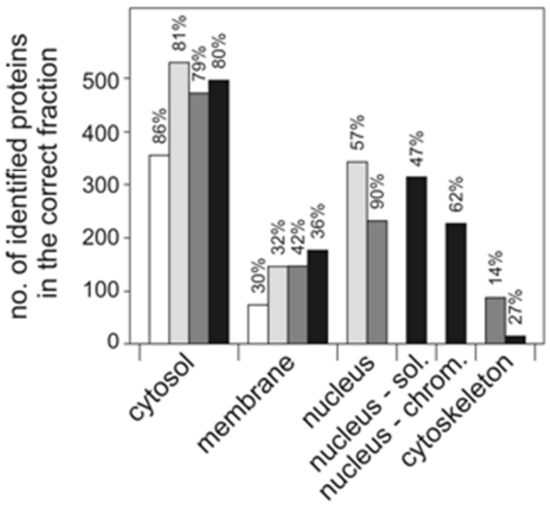


Figure 4. Evaluation of protein localization. For determination of the specificity of each method, the detected proteins in all fractions were analyzed in respect to their assumed localization according to GO terms using Scaffold 3. The bar chart shows the number of proteins identified in each fraction, which were expected to be in that cellular subfraction following Scaffold 3/GO annotations. On top of each bar the percentage of 'correctly' isolated proteins in the fractions is given (Method 1 = white bars, Method 2 = light grey bars; Method 3 = dark grey bars; Method 4 = black bars).

cytoskeletal fraction was validated as incorrect because of the incomplete GO annotations. So the GO annotations can only be used to get an overview of the subcellular localizations of a large dataset of proteins. If the localization of a distinct protein is of interest, then a literature search has to be made additionally.

3.6. Potential use of membrane proteins as markers for activation of Jurkat cells

Subcellular fractionation is an ideal tool to enrich and analyze different cellular compartments and low abundant proteins [28]. Due to the fractionation of the cells the less frequent membrane proteins, which otherwise are often covered by the numerous cytosolic proteins in MS measurement, can be identified and analyzed too. Surface proteins in the membrane are especially important for lymphocytes as they are needed for the recognition of antigens and cytokines and activation of other cells. Some of these surface proteins can be used as markers in the evaluation for different purposes. Activated lymphocytes express membrane proteins like CD25, CD69, CD71, and HLA-DR [29-32] which are absent or expressed only in low amounts on resting cells. These proteins are used as activation markers [33]. Similarly a number of known surface proteins like CD2, CD3 and CD5 were identified in the membrane fractions analyzed. In particular for CD2 and CD3 it is long known that they are involved in transmembrane signaling [34]. Despite the known marker, the analysis of the enriched membrane proteins gained by the subcellular fractionation could furthermore lead to the identification of new activation markers, when comparing the membrane proteome of resting and activated cells. Additionally, the identification and subcellular assignment of previously unknown proteins is conceivable. Newly identified membrane proteins may also be used to distinguish between the various T helper cell subpopulations and therefore assist in the process of revealing the different roles of T helper subsets.

4. Concluding remarks

The direct comparison between different methods allows an evidence-based decision on the method of choice for a specific research question. For some studies the mere separation of cytosolic and membrane proteins will be sufficient to perform subsequent analysis. Like for Western blotting method one provides a time-efficient solution of enrichment of certain proteins. When the analysis of the membrane fraction is of special interest the methods 2 or 4 might be favorable. If in the same instance also information about proteins with a nuclear localization it seems advisable to use method 4.

Acknowledgements

The work was supported by the German Research Council (Transregional Collaborative Research Centre 67, subproject Z4) and by the Helmholtz Impulse and Networking Fund through the Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE).

References

1. Nilsson, T., et al., Mass spectrometry in high-throughput proteomics: ready for the big time. Nat Methods. 7(9): p. 681-5.
2. Islinger, M., C. Eckerskorn, and A. Volkl, Free-flow electrophoresis in the proteomic era: a technique in flux. Electrophoresis. 31(11): p. 1754-63.
3. De Palma, A., et al., Extraction methods of red blood cell membrane proteins for Multidimensional Protein Identification Technology (MudPIT) analysis. J Chromatogr A. 1217(33): p. 5328-36.
4. Valot, B., S. Gianinazzi, and D.G. Eliane, Sub-cellular proteomic analysis of a Medicago truncatula root microsomal fraction. Phytochemistry, 2004. 65(12): p. 1721-32.
5. Liu, H., D. Lin, and J.R. Yates, 3rd, Multidimensional separations for protein/peptide analysis in the post-genomic era. Biotechniques, 2002. 32(4): p. 898, 900, 902 passim.
6. Kislinger, T., et al., Multidimensional protein identification technology (MudPIT): technical overview of a profiling method optimized for the comprehensive proteomic investigation of normal and diseased heart tissue. J Am Soc Mass Spectrom, 2005. 16(8): p. 1207-20.
7. Warren, C.M., et al., Sub-proteomic fractionation, iTRAQ, and OFFGEL-LC-MS/MS approaches to cardiac proteomics. J Proteomics. 73(8): p. 1551-61.
8. Michelsen, U. and J. von Hagen, Isolation of subcellular organelles and structures. Methods Enzymol, 2009. 463: p. 305-28.
9. Rio, D.C., et al., Preparation of cytoplasmic and nuclear RNA from tissue culture cells. Cold Spring Harb Protoc. 2010(6): p. pdb prot5441.
10. Sims, N.R. and M.F. Anderson, Isolation of mitochondria from rat brain using Percoll density gradient centrifugation. Nat Protoc, 2008. 3(7): p. 1228-39.
11. Kelson, T.L., J.R. Secor McVoy, and W.B. Rizzo, Human liver fatty aldehyde dehydrogenase: microsomal localization, purification, and biochemical characterization. Biochim Biophys Acta, 1997. 1335(1-2): p. 99-110.
12. Thiery, J., et al., Isolation of cytotoxic T cell and NK granules and purification of their effector proteins. Curr Protoc Cell Biol. Chapter 3: p. Unit3 37.
13. Ramsby, M.L., G.S. Makowski, and E.A. Khairallah, Differential detergent fractionation of isolated hepatocytes: biochemical, immunochemical and two-dimensional gel electrophoresis characterization of cytoskeletal and noncytoskeletal compartments. Electrophoresis, 1994. 15(2): p. 265-77.
14. Sawhney, S., R. Stubbs, and K. Hood, Reproducibility, sensitivity and compatibility of the ProteoExtract subcellular fractionation kit with saturation labeling of laser microdissected tissues. Proteomics, 2009. 9(16): p. 4087-92.
15. Churchward, M.A., et al., Enhanced detergent extraction for analysis of membrane proteomes by two-dimensional gel electrophoresis. Proteome Sci, 2005. 3(1): p. 5.
16. Solstad, T., et al., Quantitative proteome analysis of detergent-resistant membranes identifies the differential regulation of protein kinase C isoforms in apoptotic T cells. Proteomics. 10(15): p. 2758-68.
17. Wang, Y., et al., Cellular uptake of exogenous human PDCD5 protein. J Biol Chem, 2006. 281(34): p. 24803-17.
18. Cox, B. and A. Emili, Tissue subcellular fractionation and protein extraction for use in mass-spectrometry-based proteomics. Nat Protoc, 2006. 1(4): p. 1872-8.

19. Perkins, D.N., et al., Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 1999. 20(18): p. 3551-67.
20. Sirover, M.A., New nuclear functions of the glycolytic protein, glyceraldehyde-3-phosphate dehydrogenase, in mammalian cells. J Cell Biochem, 2005. 95(1): p. 45-52.
21. Tisdale, E.J., Glyceraldehyde-3-phosphate dehydrogenase is required for vesicular transport in the early secretory pathway. J Biol Chem, 2001. 276(4): p. 2480-6.
22. Glaser, P.E. and R.W. Gross, Rapid plasmenylethanolamine-selective fusion of membrane bilayers catalyzed by an isoform of glyceraldehyde-3-phosphate dehydrogenase: discrimination between glycolytic and fusogenic roles of individual isoforms. Biochemistry, 1995. 34(38): p. 12193-203.
23. Singh, R. and M.R. Green, Sequence-specific binding of transfer RNA by glyceraldehyde-3-phosphate dehydrogenase. Science, 1993. 259(5093): p. 365-8.
24. Morgenegg, G., et al., Glyceraldehyde-3-phosphate dehydrogenase is a nonhistone protein and a possible activator of transcription in neurons. J Neurochem, 1986. 47(1): p. 54-62.
25. Meyer-Siegler, K., et al., A human nuclear uracil DNA glycosylase is the 37-kDa subunit of glyceraldehyde-3-phosphate dehydrogenase. Proc Natl Acad Sci U S A, 1991. 88(19): p. 8460-4.
26. Fuchtbauer, A., et al., Actin-severing activity copurifies with phosphofructokinase. Proc Natl Acad Sci U S A, 1986. 83(24): p. 9502-6.
27. Huitorel, P. and D. Pantaloni, Bundling of microtubules by glyceraldehyde-3-phosphate dehydrogenase and its modulation by ATP. Eur J Biochem, 1985. 150(2): p. 265-9.
28. Huber, L.A., K. Pfaller, and I. Vietor, Organelle proteomics: implications for subcellular fractionation in proteomics. Circ Res, 2003. 92(9): p. 962-8.
29. Nakamura, S., et al., Human T cell activation. IV. T cell activation and proliferation via the early activation antigen EA 1. J Exp Med, 1989. 169(3): p. 677-89.
30. Ko, H.S., et al., Ia determinants on stimulated human T lymphocytes. Occurrence on mitogen- and antigen-activated T cells. J Exp Med, 1979. 150(2): p. 246-55..
31. Waldmann, T.A., The structure, function, and expression of interleukin-2 receptors on normal and malignant lymphocytes. Science, 1986. 232(4751): p. 727-32.
32. Neckers, L.M. and J. Cossman, Transferrin receptor induction in mitogen-stimulated human T lymphocytes is required for DNA synthesis and cell division and is regulated by interleukin 2. Proc Natl Acad Sci U S A, 1983. 80(11): p. 3494-8.
33. Caruso, A., et al., Flow cytometric analysis of activation markers on stimulated T cells and their correlation with cell proliferation. Cytometry, 1997. 27(1): p. 71-6.
34. Bagnasco, M., et al., Transmembrane signaling via both CD3 and CD2 human T cell surface molecules involves protein kinase-C translocation. Ric Clin Lab, 1989. 19(3): p. 221-9.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.55

Ultrasonic Enhanced Applications in Proteomics Workflows: single probe versus multiprobe

Luz Fernandez[1], Hugo. M. Santos[1], J. D. Nunes-Miranda[2], Carlos Lodeiro[1,2], José. L. Capelo*[1,2].

[1]REQUIMTE, Departamento de Química, Centro de Química Fina e Biotecnologia,Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516 Monte de Caparica, Portugal; [2]Physical Chemistry Department. Science Faculty, University of Vigo. E-32004. Ourense Spain.

Received: 12 November 2010 **Accepted:** 22 November 2010 **Available Online**: 25 November 2010

ABSTRACT

A 96-well plate-based platform in conjunction with an ultrasonic multiprobe of four tips was assessed to develop various fast proteomics workflows for gel-based proteomics. The use of such protocols reduce sample time and handling, allowing rapid processing whilst reducing the risk of contamination. The procedure reduces the time to indentify proteins separated by gel electrophoresis to just 8 min/each. In addition, the ultrasonic multiprobe was compared with the single probe as a tool to obtain high sample throughput in proteomics workflows entailing identification and/or quantification of proteins using mass-spectrometry based approaches. The 18-O labeling-based method was used to study the type of peptides extracted from the gels when the extraction was done with the aid of ultrasonic energy. The assessment was done in ten standard proteins separated by gel elecrophoresis. Two proteins obtained from *D. desulfuricans,* and from *Cyprinus carpio*, Split-Soret cytochrome c, and Vitellogenin respectively, were also indentified as a further proof-of-the concept.

Keywords: Ultrasonic, MALDI, Vitellogenin, ^{18}O, Inverse labeling.

1. Introduction

Ultrasonication has been recently appointed as a powerful tool in mass spectrometry-based proteomics workflows for protein identification [1-7]. Ultrasonic energy can be used to enhance from hours to minutes protein denaturation, protein reduction, protein alkylation and protein digestion, the four mais steps of any common procedure nowadays used for protein identification relaying on mass spectrometry. Furthermore, ultrasonic energy can also be used to speed protocols relaying on ^{18}O isotopic labeling, which is a widely used method to tracking changes in protein level expression as well as in sequencing of peptides by mass spectrometry-based techniques [6,7]. As a matter of fact ultrasonic energy has been recently integrated in rapid sample processing for ^{18}O-LC-MS-based quantitative proteomics [8].

Ultrasonic-based high throughput sample treatment for proteomics was recently reported for the treatment of liquid samples by joining a 96-well plate and an ultrasonic multiprobe. The present work shows a step forward of this protocol by applying it to proteins separated by gel-based approaches and also using it to study in an 18O labeling-based method the type of peptides extracted from the gels when the extraction is done with the aid of ultrasonic energy. The study was done through the identification of 10 standard proteins and two proteins obtained from *D. desulfuricans,* and from Cyprinus carpio, Split-Soret cytochrome c, and Vitellogenin respectively.

2. Material and methods

2.1 *Apparatus*

Protein digestion was done in a 96-well plate (Digilab-Genomic Solutions, USA). A vacuum concentrator centrifuge from UniEquip (Martinsried, Germany) model UNIVAPO 100H with a refrigerated aspirator vacuum pump model Uni-

***Corresponding author:** José. L. Capelo. Tel.: +34 610 835 903; Email Address: jlcapelom@uvigo.es; URL: http://www.bioscopegroup.com (J.L. Capelo).

jet II was used for (i) sample drying and (ii) sample preconcentration. A minicentrifuge, model Spectrafuge-mini, from Labnet (Madrid, Spain), and a minicentrifuge-vortex, model Sky Line, from ELMI (Riga, Latvia) were used throughout the sample treatment, when necessary. A Simplicity™ 185 from Millipore (Milan, Italy) was used to obtain Milli-Q water throughout the experiments.

2.2 *Ultrasonic devices*

(i) Ultrasonic probe, model UP 100H (dr. Hielscher, Teltow, Switzerland, 200 W, 30 kHz ultrasonic frequency, 0.5 mm of diameter probe).

(ii) Ultrasonic multi-probe from Branson Ultrasonics Corporation (USA), model SLPe (150 W, 40 kHz ultrasonic frequency, 1 mm diameter probe). The ultrasonic generator SLPe is equipped with a multi-probe detachable horn (model 4c15), with four tips for simultaneous ultrasonication of four samples and it was used in conjunction with a 96-well plate, as it is depicted in video 1 of supporting information.

2.3 *Standards and reagents*

The following protein standards were used: α-lactalbumin from bovine milk (≥85%), BSA (>97%) and carbonic anhydrase (>93%) from Sigma (Steinheim, Germany), albumin from hen white (>95%) from Fluka (Buchs, Switzerland). Chymotrypsinogen A, catalase bovine and aldolase from rabbit were standards for gel filtration calibration kit high molecular weight from Amersham Biosciences (Piscataway, USA).

Low molecular weight standard protein mixture of glycogen phosphorylase b, bovine serum albumin, BSA, ovalbumin, carbonic anhydrase, trypsin inhibitor and α-lactalbumin were purchased from Amersham Biosciences (Piscataway, USA). Thyroglobulin and Lactate dehydrogenase were purchased from Amersham Biosciences (Piscataway, USA)

Carp vitellogenin standard was purchased from Biosense Laboratories (Bergen, Norway).

Trypsin enzyme, sequencing grade was purchased from Sigma. All materials were used without further purification. α-cyano-4-hydroxycinnamic acid (α-CHCA) puriss for MALDI–MS from Fluka was used as MALDI matrix. ProteoMass™ Peptide MALDI-MS Calibration Kit (MSCAL2) from Sigma was used as mass calibration standard for MALDI-TOF-MS.

The following reagents were used for protein depletion: sodium chloride puriss. p.a. and magnesium chloride hexahydrate puriss. p.a. were purchased from Fluka; ethylenediaminetetraacetic acid disodium salt dehydrate puriss. p.a. was from Riedle-de Haën (Seelze, Germany).

The following reagents were used for protein digestion: acetonitrile, iodoacetamide (IAA) and DL-dithiothreitol (DTT) (99%) were purchased from Sigma; formic acid and ammonium bicarbonate (>99.5%) were from Fluka; trifluoroacetic acid (TFA, 99%) were from Riedel-de-Haën (Seelze, Germany); and urea (99%) was from Panreac (Barcelona, Spain).

2.4 *Sample treatment*

2.4.1. Protein separation by 1D-SDS-PAGE

Amounts of protein ranging from 0.5 to 3.7 μg were dissolved in 5 μL of water plus 5 μL of sample buffer (5 mL of 0.5 M Tris-Base + 8 mL of 10 % SDS + 1 mL of β-mercaptoethanol + 2 mL of glycerol + 4 mg of bromophenol blue in a final volume of 20 mL in water) for analysis by sodium dodecyl sulphate polyacrilamide gel electrophoresis (SDS-PAGE) (10% 0.5 mm thickness). After gel electrophoresis (65 min, 120 V, 400 mA), the gel was stained with Coomassie blue R-250 and destained in order to visualize the proteins bands.

2.4.2. In-gel sample treatments

(i) Overnight method. For in-gel digestion optimization, 2.1 μg of BSA and 2.9 μg of α-lactalbumin were loaded onto 10% SDS-PAGE gels. Coomassie Blue-stained protein bands were excised from the gels, cut into pieces and subjected to digestion. Excised gel bands were then washed with water (3 times with agitation/centrifugation, 10min each), and dehydrated with acetonitrile (2 times, 3 min each + 1 time, 20 min with agitation/centrifugation) and dried in a vacuum centrifuge. Gel pieces were further rehydrated with 10 mM of DTT in 25 mM ammonium bicarbonate buffer and incubated 10 min at 60 ºC for protein reduction. Then DTT solution was replaced by IAA 55 mM in 25 mM ammonium bicarbonate buffer and incubated in the dark and room temperature by 35 min. After protein reduction and alkylation gel pieces were dried and rehydrated in ice bath in a 0.025 μg/μL solution of trypsin in 12.5 mM ammonium bicarbonate buffer, to a final volume of 25 μL, during 1 h.

After the rehydration step, samples were digested overnight at 37 °C. Next, trypsin activity was stopped by the addition of 20 μL of 5% formic acid. The supernatant was withdrawn and retained, and the peptides were extracted from the gel pieces by adding 50-100 μL of a mixture of acetonitrile/TFA (500 μL H_2O+500 μL AC+1 μL TFA) and incubating them for 15 min at 37 ºC in a shaker. This extraction was done twice. Then, all supernatants were combined and evaporated to dryness in a vacuum concentrator centrifuge and finally the dried peptides obtained were reconstituted with 10 μL of 0.3% v/v formic acid.

(ii) Accelerated method. In this method, the protocol described above and referred as "overnight method" was followed but (i) washing steps (ii) alkylation, (iii) reduction, (iv) digestion of gel bands and (v) extraction peptides were done in 8 min (2 min each washing step, 30% ultrasonic amplitude, total of 3), 5 min (30% ultrasonic amplitude), 5 min (30% ultrasonic amplitude), 4 min (25% ultrasonic amplitude) and 8 min (two extraction steps, 2 min each) respectively, using ultrasonication at 30 kHz, with a single probe or 40 kHz with the four tip multiprobe.

2.5 *Case studies*

2.5.1. Desulfovibrio desulfuricans ATCC27774

Desulfovibrio desulfuricans ATCC27774 cells were cultured in sulfate-lactate medium. Cells were collected by centrifugation (8000 × g during 15 min at 4 ºC), resuspended in 10 mM Tris-HCl buffer and ruptured in a French press at 9000 psi. After centrifugation (10000 × g, 45 min) and ultracentrifugation (180000 × g, 60 min) the supernatant was dialyzed against 10 mM Tris-HCl buffer. Both proteins were isolated from the soluble extract using chromatographic columns (anionic exchange, Hydroxyapatite column and molecular exclusion chromatography). The purity of the proteins was evaluated by SDS-PAGE and UV-visible spectroscopy. All purification procedures were performed under aerobic conditions at 4 ºC and pH 7.6. Split-soret cytochrome c from *Desulfovibrio desulfuricans* ATCC27774 was in gel digested according to the accelerated method described in 2.5.2. Protein identification was done using the PMF procedure by MALDI-TOF-MS.

2.5.2 Plasmatic vitellogenin from Cyprinus carpio

80 µL of plasma from *Cyprinus carpio* was diluted to 100 µL with cold PBS (Phosphate Buffer Solution). 300 µL of -20 ºC cold acetone were added into the diluted plasma solution and kept overnight on ice. The sample was centrifuged at 10000 g, 4 ºC for 30 min. The supernatant was removed and the pellet was suspended in 100 µL of buffer (10 mM Tris-HCl pH 7.4; 2% of SDS; 1% of β-mercaptoethanol) [21]. Amounts of delipidated plasma (5 µL) were mixed with 5 µL of sample buffer (5 mL of 0.5 M Tris-Base + 8 mL of 10 % SDS + 1 mL of β-mercaptoethanol + 2 mL of glycerol + 4 mg of bromophenol blue in a final volume of 20 mL in water) for analysis by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) (7.5% 0.5 mm thickness). After gel electrophoresis (65 min, 120 V, 400 mA), the gel was stained with Coomassie blue R-250 and destained in order to visualize the proteins bands. Vitellogenin was in gel digested according to the accelerated method described in 2.5.2. Protein identification was done using the PMF procedure by MALDI-TOF-MS.

2.6 *Inverse ^{18}O labeling of peptides.*

Protein BSA was used to study the effect of ultrasonication in the release of peptides from gels. BSA was separated by 1D-PAGE, and then submitted to the protocol described in section 2.5.2 (ii). Once the protein was digested, gel pieces were removed and the solutions containing the peptides were dried in a speed vacuum. Then, peptides obtained were recomposed in (i) normal water or (ii) ^{18}O water and then both methods were compared following the inverse ^{18}O labeling protocol as described by Wang et al [10].

2.7 *MALDI-TOF-MS analysis*

A MALDI-TOF-MS model Voyager DE-PRO Biospectrometry Workstation equipped with a nitrogen laser radiating at 337 nm from Applied Biosystems (Foster City, USA), was used to obtain the PMF. MALDI mass spectra were acquired as recommended by the manufacturer and treated with the Data Explorer™ software version 4 series. Prior to MALDI-TOF-MS analysis, the sample was mixed with the matrix solution. α-CHCA matrix was used throughout this work and was prepared as follows: 10 mg of α-CHCA was dissolved in 1 mL of Milli-Q water/acetonitrile/TFA (1mL+1mL+2µL). Then, 2 µL of the aforementioned matrix solution was mixed with 2 µL of sample and the mixture was shaken in a vortex for 30 s. Finally, 1 µl of the sample/matrix mixture was spotted on a well of a MALDI-TOF-MS sample plate and was allowed to dry.

Measurements were done in the reflector positive ion mode, with a 20 kV accelerating voltage, 75.1 % grid voltage, 0.002 % guide wire and a delay time of 100 ns. Two close external calibrations were performed with the monoisotopic peaks of the Bradykinin, Angiotensin II, P14R and ACTH peptide fragments (m/z: 757.3997, 1046.5423, 1533.8582 and 2465.1989, respectively). Monoisotopic peaks were manually selected from each of the spectra obtained. Mass spectral analysis for each sample was based on the average of 500 laser shots. Peptide mass fingerprints were searched with the MASCOT [http://www.matrixscience.com/search_form_select.htmL] search engine with the following parameters: (i) SwissProt. 2006 Database; (ii) molecular weight (MW) of protein: all; (iii) one missed cleavage; (iv) fixed modifications: carbamidomethylation (C); (v) variable modifications: oxidation (M); (vi) peptide tolerance up to 150 ppm. A match was considered successful when the protein identification score is located out of the random region and the protein analyzed scores first.

3. Results and Discussion

3.1. *96 well plate method for proteins separated by gel electrophoresis*

The protein concentration loaded onto the gel was 2,1 µg/µL for BSA and 2.9 µg/µL for α-lactalbumin. Results, including the values setting for each variable, are presented in Fig. 1. As may be seen, ultrasonic amplitude was found a critical parameter. For BSA protein, when the lower amplitude was used in the digestion step (10%), longer treatment times were required to obtain good results, whilst the highest amplitude used (50%) degrades the gel in such a way that protein identification by MALDI was not possible. In the case of the α-lactalbumin this problem was even worst since using the lower amplitude in the digestion step no identification was possible at all in the range time assessed (60s-240s), whilst for the highest amplitude the identification was only possible in times comprised between 120 and 240 s. These results could be explained on the basis of recent research in ultrasonic applications in medicine and drug delivery that has estimated the pressure at the tip of the jet generated by bubble collapse (cavitation phenomena) around 60 MPa [11]. This is high enough to penetrate small pores, such as the ones

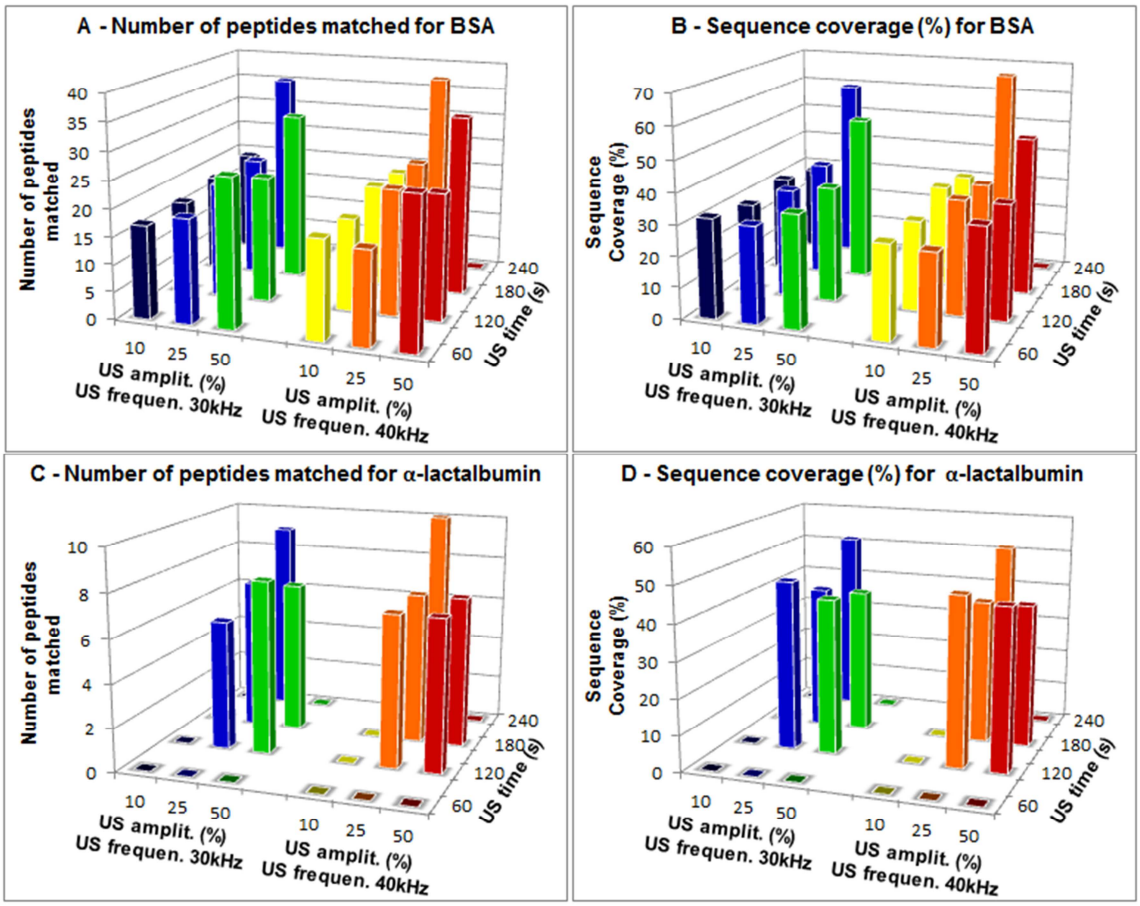


Figure 1. Number of peptides matched and sequence coverage for BSA and α-lactalbumin as a function of time, amplitude and frequency of sonication. Proteins were separated by Gel electrophoresis. Peptides matched and sequence coverage for the overnight method was 42±6 and 70±4 respectively for BSA and 11±2 and 51±1 for α-lactalbumin.

present in the gels in which proteins are separated. Hence, liquid jets may act as microsyringes, delivering the enzyme to a region of interest. However, the ultrasonic energy needed to do such delivering must be reached, and we hypothesized that this is the reason why the lower amplitude studied did not perform well, because it is not powerful enough to deliver the enzyme inside the gel, as it was the case for α-lactalbumin. As the amplitude and time is increased, the enzymatic cleavage performs better for both proteins, until a maximum is reached. Then, the gel is degraded and the protein identification becomes difficult or impossible, as it was the case for both proteins for the highest amplitude and time studied. Gel degradation can be explained on basis on the fact that when a cavitation bubble collapses near the surface of a solid sample particle, micro-jets of solvent propagate toward the surface at velocities greater than 100 m s^{-1}, causing pitting and mechanical erosion of the solid surface, thus leading to particle rupture (i.e., disruption) [12]. Although this process could be favorably used to enhance peptide release from the gel, at some point it becomes a problem interfering the subsequent analysis by mass spectrometry or even blocking HPLC columns. However, it is noteworthy that the multi-probe and the 96-well plate performed well for gel-based protein separations, once the process has been optimized. In addition, it was not found differences between the two frequencies studied. Furthermore, the best results obtained with the multi-probe were not different from the ones obtained using the overnight protocol.

To complete the study a set of 8 further proteins was mixed and separated using gel electrophoresis to proceed to protein identification through the ultrasonic method. Results showed in Table 1 demonstrate the successful accomplishment between the 96-well plate and the multi-probe since all proteins were identified with similar protein coverage and number of peptides matched than using the overnight protocol.

A comparison of the total time and number of steps involved in the handling of the five sample treatments reported in this manuscript is presented in Table 2. As may be seen, the ultrasonic method with multiprobe allows for the treatment of 4 samples in 30 minutes. This number can be exponentially increased using the new generation of multiprobes that allow treating 96 samples at once [11]. Hence, this method has a great potential for clinical purposes, where a high number of samples are generally handled daily.

3.2 *Effect of ultrasonication in the release of peptides from gels.*

To investigate the role of ultrasonication in the release of

Table 1. Number of peptides matched and protein sequence coverage for in gel-protein digestion by the overnight method and accelerated method.

Protein	theor. M_r (kDa)	overnight method (n = 2)			accelerated method (n = 2)					
		Mascot score	Sequence coverage (%)	No. of peptides matched	Mascot score		Sequence coverage (%)		No. of peptides matched	
					30 kHz	40 kHz	30 kHz	40 kHz	30 kHz	40 kHz
α-Lactalbumin	16.7	130±5	51±1	11±2	110±2	104±5	50±2	51±0	9±1	10±0
Trypsin inhibitor	24.3	87±3	51±1	18±0	69±3	64±5	28±2	27±1	7±1	7±0
Carbonic anhydrase	29.1	115±3	69±0	20±0	240±2	238±3	71±2	64±0	19±0	18±1
Ovalbumin	43.2	100±2	49±6	20±1	82±2	84±5	50±2	44±4	16±2	15±0
BSA	71.2	152±4	70±4	42±6	220±1	222±3	64±2	66±5	40±3	37±1
Aldolase rabit	39.8	152±6	72±6	21±2	109±3	103±4	68±1	68±0	20±1	19±0
Catalase bovine	60.1	98±4	43±2	26±2	218±2	220±0	37±4	43±0	22±1	24±0
Phosphorylase b	97.7	92±5	66±1	64±1	298±1	305±3	48±3	52±2	40±2	38±1
Thyroglobulin subunit*	310.0	194±8	28±1	76±6	188±2	194±6	28±1	25±7	80±2	81±1
Lactate dehydrogenase subunit*	36.9	79±2	46±4	21±1	83±1	182±4	46±2	49±0	20±1	21±0
Split-Soret cytochrome c *D. desulfuricans*	27.8	128±4	51±4	11±1	100±4	96±6	40±5	37±0	10±2	10±1
Vitellogenin *Cyprinus carpio*	148.8	176±6	42±4	49±1	150±3	141±6	42±1	41±1	50±2	51±1

* HMW-Native standard under denaturant conditions.

peptides from gels, a set of experiments using the ^{18}O isotopic labeling of the protein BSA was done to compare, the type of peptides released using the overnight, ON, or the ultrasonic, US, treatments [6,7]. The same amount of BSA was loaded onto a gel, and the protein was in-gel digested, using the overnight or the ultrasonic protocol. The supernatants were withdrawn from the tubes and dried down in a vacuum centrifuge. Then samples were recomposed in ^{16}O or in ^{18}O water. Since under conventional conditions, the yield of peptides extracted from a gel is protein-dependant and it varies for different peptides that originate from the same protein, we decide to use the inverse labeling as described by Wang et al [10] to ensure a correct identification of peptides extracted and unambiguous assessment of differential peptide extraction. To avoid any biased yields own to the sample treatment, all steps were exactly the same for the overnight or ultrasonic protocols, with the exception of the protein digestion step. Fig. 2 shows the inverse labeling method for the unambiguous identification of peptides used. The result of this set of experiments are shown in Table 3, where may be seen that a total of 40 peptides were identified as BSA peptides. Interestingly, from those 40 peptides, 15 peptides were labeled either in the direct or in the reverse process, whilst 21 peptides from the total, were also identified but they were not labeled. This put forwards that after sample recomposition not all peptides are labeled. This result is in agreement with data published by other authors [13]. From this 21 non labeled peptides, 6 were observed in the direct and reverse method, suggesting that they are produced regardless of the digestion method used; 9 were found in the mixture US^{16}O / ON^{18}O and six were found in the mixture US^{18}O / ON^{16}O. Therefore it can be concluded that some peptides are preferentially formed as a function of the digestion method. Thus, peptides (GACLLPK)H^+; (LCVLHEKTPVSEK)H^+; (MPCTEDYLSLILNR)H^+ are formed in the overnight digestion, whilst peptides (VLASSAR)H^+; (ALKAWSVAR)H^+; (HLVDEPQNLIK)H^+ are formed in the ultrasonic method. On the overall, both methods produced the same number of total peptides, c.a. 40, confirming that both methodologies work well for protein digestion. Another interesting finding is that protein modifications as consequence of the heating/cavitation caused by the ultrasonic energy were not

Table 2. Comparison of handling and time consumed for the five methods studied with the 96 well plate ultrasonic method.

		Denaturation	Reduction	Alkylation	Digestion	Desalting	Total steps	Total time
Proteins in gel	Overnight	5 min heating before electrophoresis	10 min US	35 min US	12 hours US	no	25*	24 h
	Ultrasonic	5 min heating before electrophoresis	5 min	5 min	4 min	no	20**	30 min

* Including: electrophoresis, band excision, 12 steps for gel washing, 10×3 min each and 2×20 min each, trypsin incubation on ice before digestion, peptides extraction (2 extractions) and evaporated to dryness (3 times).

** Including: electrophoresis, band excision, 8 steps for gel washing, 2×2 min each, digestion and peptides extraction and evaporated to dryness (3 times). With the in gel ultrasonic digestion no incubation of trypsin on ice is needed.

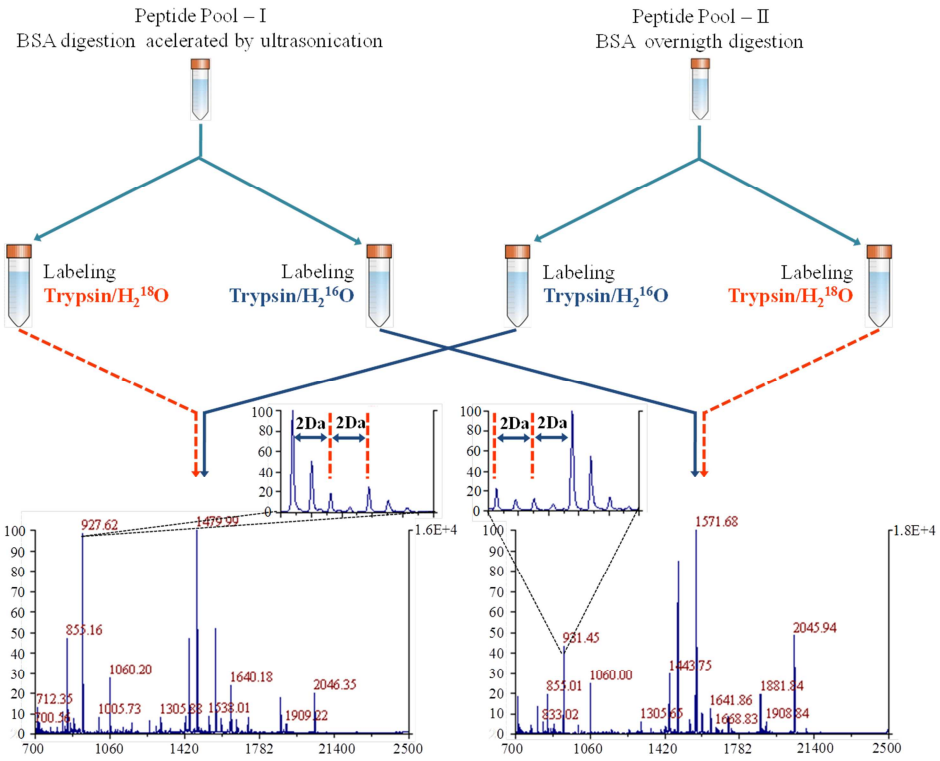


Figure 2. The inverse labeling method for the unambiguous identification of peptides released from gels using the overnight or the ultrasonic digestion protocol.

Table 3. Results from the BSA ^{18}O-inverse labeling experiments. All peptides were manually verified. (n=2)

	Nº. of peptides	Peptide fragment
Direct & reverse Labeled peptides US^{16}O / ON^{18}O and US^{18}O / ON^{16}O	15±2	(SEIAHR)H$^+$; (YLYEIAR)H$^+$; (LVNELTEFAK)H$^+$; (HPEYAVSVLLR)H$^+$; (SLHTLFGDELCK)H$^+$; (RHPEYAVSVLLR)H$^+$; (YICDNQDTISSK)H$^+$; (TCVADESHAGCEK)H$^+$; (LGEYGFQNALIVR) H$^+$; (HPEYAVSVLLRLAK)H$^+$; (KVPQVSTPTLVEVSR) H$^+$; (MPCTEDYLSLILNR)H$^+$; (RPCFSALTPDETYVPK) H$^+$; (LFTFHADICTLPDTEK)H$^+$; (RHPYFYAPELLYYANK)H$^+$
Direct & reverse non-labeled peptides US^{16}O / ON^{18}O and US^{18}O / ON^{16}O	6±1	(CASIQK)H$^+$; (TPVSEKVTK)H$^+$; (CCTESLVNR)H$^+$; (DTHKSEIAHR)H$^+$; (FKDLGEEHFK)H$^+$; (TVMENFVAFVDK)H$^+$
Labeled peptides from mixture US^{16}O / ON^{18}O	3±1	(GACLLPK)H$^+$; (LCVLHEKTPVSEK)H$^+$; (MPCTEDYLSLILNR)H$^+$ Oxidation (M)
Non-labeled peptides from mixture US^{16}O / ON^{18}O	9±1	(ALKAWSVAR)H$^+$; (HLVDEPQNLIK)H$^+$; (TVMENFVAFVDK)H$^+$ Oxidation (M); (VTKCCTESLVNR)H$^+$; (ETYGDMADCCEK)H$^+$; (LKHLVDEPQNLIK)H$^+$; (DDPHACYSTVFDK)H$^+$; (AEFVEVTKLVTDLTK)H$^+$; (YNGVFQECCQAEDK)H$^+$; (HPYFYAPELLYYANK)H$^+$;
Labeled peptides from mixture US^{18}O / ON^{16}O	3±0	(VLASSAR)H$^+$; (ALKAWSVAR)H$^+$; (HLVDEPQNLIK)H$^+$
Non-labeled peptides from mixture US^{18}O / ON^{16}O	6±1	(LVTDLTK)H$^+$; (LSQKFPK)H$^+$; (IETMREK)H$^+$; (SEIAHRFK)H$^+$; (NECFLSHK)H$^+$; (CCTKPESER)H$^+$; (LKECCDKPLLEK)H$^+$

found. For instance, no extra oxidations or carbamydomethcarbamydomethylations were detected.

These results confirm the utility of ^{18}O labeling in relative proteomics discovering and confirms the usefulness of the combination of ultrasonication and a 96-well plate for proteomics applications.

4. Future prospects

The speed, simplicity, high throughput and number of potential proteomics applications that can be developed jointing 96-well plate and ultrasonic multiprobes, it makes of this combination an ideal tool for robotic platforms. As an example, if a multiprobe of 96 tips was acomplisehd with a 96

well plate, the time need to identified a protein separated by gel electrophoresis could be reduced to just 20 s. Therefore developments in this area of research are anticipated.

5. Conclusions and perspectives

It has been proven that the combination of a 96-well plate and an ultrasonic multi-probe is a powerful tool in sample treatment for proteomics, allowing high sample throughput and a potentially enormous number of different proteomics applications. The huge variety of protocols that can be used with this accomplishment has it been demonstrated through different proteomics sample treatments for protein identification and 18-O based labelling. Sample preparation steps including destaining, washing, reduction & alkylation, digestion, spotting on MALDI targets or transfer to LC/MS input plates can be combined on a single automated platform making use of ultrasonication. This allows for rapid processing, minimizes the risk of contamination and therefore reduces the chance of application errors and improves the quality of data. The results showed that for protein digestion low or high ultrasonic amplitudes must be avoided when using a 96 well plate and an ultrasonic multiprobe. We have not found differences in performance for protein identification when using ultrasonic amplitudes of 30 (single probe) or 40 kHz (multiprove).

We have demonstrated that using the direct and reverse ^{18}O labeling the effectiveness of different procedures for in-gel protein digestion can be compared in terms of number and type of peptides produced. In fact our findings showed a similar number of peptides obtained by either the overnight or the ultrasonic method. However, some peptides were preferentially formed for each digestion protocol.

Acknowledgements

H. M. Santos and R. Carreira acknowledge the doctoral grants SRFH/BD/38509/2007 and SFRH/BD/28563/2006, respectively, from FCT (Science and Technological Foundation) from Portugal. The research findings here reported are protected by international laws under patent pendings PCT/IB2006/052314 and PT 103 303. FCT is acknowledged for financial support under the project POCI/QUI/55519/2004 FCT-FEDER and PTDC/QUI-QUI/099907/2008. J. L. Capelo and C. Lodeiro acknowledge the Isidro Parga Pondal Program from Xunta de Galicia, Spain. Gabriela Ribas and Mario Diniz are acknowledge for kindly provide us with *D. desulfuricans* ATCC27774 and *C. carpio* samples, respectively.

References

1. D. Lopez-Ferrer, B. Canas, J. Vazquez, C. Lodeiro, R. Rial-Otero, I. Moura, J.L. Capelo, Sample treatment for protein identification by mass spectrometry-based techniques. Trac-Trends Anal. Chem. 25 (2006) 996-1005.
2. R.J. Carreira, R. Rial-Otero, D. Lopez-Ferrer, C. Lodeiro, J.L. Capelo, Ultrasonic energy as a new tool for fast isotopic O-18 labeling of proteins for mass spectrometry-based techniques: Preliminary results, Talanta, 76 (2008) 400-406.
3. D. Lopez-Ferrer, J.L. Capelo, J. Vazquez, Ultra-fast trypsin digestion of proteins by high intensity focused ultrasound. J. Proteome Res. 4 (2005) 1569-1574.
4. H.M. Santos, C. Mota, C. Lodeiro, I. Moura, I. Isaac, J.L. Capelo, An improved clean sonoreactor-based method for protein identification by mass spectrometry-based techniques, Talanta, 77 (2008) 870-875.
5. H.M. Santos, J.L. Capelo, Trends in ultrasonic-based equipment for analytical sample treatment. Talanta 73 (2007) 795-802.
6. R.J. Carreira, R. Rial-Otero, D. Lopez-Ferrer, C. Lodeiro, J.L. Capelo, Ultrasonic energy as a new tool for fast isotopic O-18 labeling of proteins for mass spectrometry-based techniques: Preliminary results, Talanta, 76 (2008) 400-406.
7. Santos, H.M., Glez-Peña, D., Reboiro-Jato, M., Fdez-Riverola, F., Diniz, M.S., Lodeiro, C., Capelo-Martínez, J.-L., Novel 18O inverse labeling-based workflow for accurate bottom-up mass spectrometry quantification of proteins separated by gel electrophoresis, Electrophoresis, 31 (2010) 3407-3419.
8. D. Lopez-Ferrer, T.H. Heibeck, K. Petritis, K.K. Hixson, W. Qian, M.E. Monroe, A. Mayampurath, R.J. Moore, M.E. Belov, D.G. Camp II, R.D. Smith, J. Proteome Res. 7 (2008) 3860-3865.
9. Ultrasonic multiprobe as a new tool to overcome the bottleneck of throughput in workflows for protein identification relaying on ultrasonic energy, Santos, H.M., Carreira, R., Diniz, M.S., Rivas, M.G., Lodeiro, C., Moura, J.J.G., Capelo, J.L., Talanta 81 (2010) 55-62
10. Y.K. Wang, Z.X. Ma, D.F. Quinn, W.F. Emil, Inverse O-18 labeling mass spectrometry for the rapid identification of marker/target proteins, Anal. Chem. 72 (2001) 3742-3750.
11. M. Postema, A. Wamel, C. T. Lancee, N. Jong, Ultrasound-induced encapsulated microbubble phenomena. Ultrasound Med. Biol. 30 (2004) 827-840.
12. G. Wibetoe, D. T. Takuwa, W. Lund, G. Sawula, Coulter particle analysis used for studying the effect of sample treatment in slurry sampling electrothermal atomic absorption spectrometry. Fresenius J. Anal. Chem. 363 (1999) 46-54.
13. X. Yao, C. Afonso, C. Fenselau, Dissection of proteolytic O-18 labeling: Endoprotease-catalyzed O-16-to-O-18 exchange of truncated peptide substrates, J. Proteome Res. 2 (2003) 147-152.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.44

Proteomic identification of plasma signatures in type 2 diabetic nephropathy

Peir-Haur Hung[1,a], Ying-Chieh Lu[2,a], Yi-Wen Chen[2], Hsiu-Chuan Chou[3], Ping-Chiang Lyu[2], Ying-Ray Lee*[,1], Hong-Lin Chan*[2].

[1]Department of [1]Department of Medical Research, Chiayi Christian Hospital, Chiayi, Taiwan; [2] Institute of Bioinformatics and Structural Biology and Department of Medical Science, National Tsing Hua University, Hsinchu, Taiwan; [3] Department of Applied Science, National Hsinchu University of Education, Hsinchu, Taiwan; [a]These authors contribute equally to this work.

Received: 25 October 2010 **Accepted:** 23 November 2010 **Available Online**: 29 November 2010

ABSTRACT

Recent advances in quantitative proteomics have offered opportunities to discover plasma proteins as biomarkers for tracking the progression and for understanding the molecular mechanisms of diabetes. We used quantitative proteomic analysis to identify novel biomarkers of nephropathy in plasma from type 2 diabetic patients. Plasma samples were analyzed by fluorescence two-dimensional differential gel electrophoresis (2D-DIGE), and differentially expressed proteins identification was performed by matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS). Proteomics analysis of the plasma proteome in type 2 diabetes mellitus with nephropathy identified 34 protein spots representing 31 unique proteins. These proteins mainly belonged to metabolic (such as 5'-AMP-activated protein kinase subunit beta-1) and growth regulatory (such as LIM homeobox protein 6) proteins. Additionally, our quantitative proteomic approach has identified numerous previous reported plasma markers of type 2 diabetes mellitus such as apolipoprotein A-I and ficolin-3. On the contrary, we have presented several putative type 2 diabetes mellitus biomarkers including calpain-7 and choline/ethanolamine kinase which have not been reported and may be associated with the progression and development of the disease. The potential of utilizing these markers for screening and treating type 2 diabetes mellitus warrants further investigation. Collectively, our results show that the proteins identified in this study may constitute potential biomarkers for the diagnosis of type 2 diabetics with nephropathy.

Keywords: Proteomics; Plasma biomarker; Type2 diabetic nephropathy.

1. Introduction

Diabetes mellitus is one of the most common metabolic diseases in the world, in which more than 90% diabetes mellitus patients are grouped to type 2 diabetes mellitus [1]. Type 2 diabetes mellitus is characterized by hyperglycemia due to defects of multiple organs which can not accurately measure the serum glucose level and uptake serum glucose. Additionally, type 2 diabetes mellitus is a chronic disease, which typically covers several decades from high serum glucose to significant diseases. The serum glucose level just reflects the outcome of multiple physiological disorders. Accordingly, numerous studies have been made to identify biomarkers to monitor the progression of type 2 diabetes mellitus [2-4]. However, most of these studies were concentrated on the genetic defects in gene expression level. Relative rare reports were given attention to protein level changes in response to the progression of diabetes. For this, proteomic strategy might provide solutions for identification of large set of the proteins in tissues or in bio-fluid associated with type 2 diabetes mellitus.

Two-dimensional gel electrophoresis (2-DE) and MALDI-TOF MS has been widely used for profiling plasma proteins and some of the nonionic and zwitterionic detergents such as thiourea and CHAPS have been introduced to increase the solubility of the plasma proteins. In addition, a significant improvement of gel-based analysis of protein quantifications and detections is the introduction of 2D-DIGE. 2D-DIGE is able to co-detect numerous samples in the same 2-DE to minimize gel-to-gel variation and compare the protein features across different gels by means of an internal fluorescent standard. This innovative technology relies on the pre-

***Corresponding author :** Ying-Ray Lee, Email Address: yingray.lee@gmail.com; Hong-Lin Chan, Email Address: hlchan@life.nthu.edu.tw.

labeling of protein samples before electrophoresis with fluorescent dyes Cy2, Cy3 and Cy5 each exhibiting a distinct fluorescent wavelength to allow multiple experimental samples to include an internal standard. Thus, the samples can be simultaneously separated in one gel. The internal standard, which is a pool of an equal amount of the experimental protein samples, can facilitate the data accuracy in normalization and increase statistical confidence in relative quantitation across gels [5-10].

In order to examine differentially expressed levels of plasma proteins associated with type 2 diabetes mellitus, a proteomics-based approach was performed involving immunodepletion of high abundance proteins, 2D-DIGE analysis and subsequent MALDI-TOF MS analysis to obtain a panel of plasma proteins found to be differentially expressed between type 2 diabetes mellitus patients and healthy donors.

2. Material and methods

2.1 *Chemicals and Reagents*

Generic chemicals and albumin and IgG depletion kit were purchased from Sigma-Aldrich (St. Louis, USA), while reagents for 2D-DIGE were purchased from GE Healthcare (Uppsala, Sweden). All the chemicals and biochemicals used in this study were of analytical grade.

2.2 *Plasma sample collection and purification*

From Jan 2009 to Dec 2009, eleven donors in a single center (Chiayi Christian Hospital, Chiayi, Taiwan) were enrolled in the study. Those included in the study were divided into type 2 diabetes mellitus patients (n=6) and healthy donors group (n=5). The criteria to assess the presence of type 2 diabetes mellitus were based on the guidelines proposed by the World Health Organization. All type 2 diabetic patients had typical diabetic symptoms along with a single fasting plasma glucose level of 7 mM or 2 hr postglucose or casual postprandial plasma glucose level of 11.1 mM. Healthy individuals with their fasting blood glucose below 5.5 mM were selected as controls. This study was approved by the Institutional Research Board and carried out according to the Helsinki Declaration Principles. Written informed consent was collected from all participating subjects.

To improve the performance of proteomic analysis of the plasma samples, the albumin and immunoglobulin G in the collected plasma samples were depleted using an albumin and IgG removal kit (Sigma, St. Louis, USA) in according with the manufacturer's instructions. The depleted plasma samples were precipitated by adding 1 volume of 100% (w/v) TCA to 4 volumes of sample and incubated for 10 min. at 4ºC. The precipitated protein was then recovered by centrifugation at 13000 rpm for 10 min. (at -20°C), and the resulting pellet was washed twice with ice-cold acetone. Air-dried pellets were resuspended in 2-DE lysis containing 4% w/v CHAPS, 7M urea, 2M thiourea, 10mM Tris-HCl, pH8.3, 1mM EDTA.

2.3 *Sample preparation for 2D-DIGE and gel image analysis*

The plasma protein pellets were dissolved in 2-DE lysis buffer and protein concentrations were determined using coomassie protein assay reagent (BioRad). Before performing 2D-DIGE, protein samples were labeled with N-hydroxy succinimidyl ester-derivatives of the cyanine dyes Cy2, Cy3 and Cy5 following the protocol described previously [11,12]. Briefly, 150 μg of protein sample was minimally labeled with 375 pmol of either Cy3 or Cy5 for comparison on the same 2-DE. To facilitate image matching and cross-gel statistical comparison, a pool of all samples was also prepared and labeled with Cy2 at a molar ratio of 2.5 pmol Cy2 per μg of protein as an internal standard for all gels. Thus, the triplicate samples and the internal standard could be run and quantify on multiple 2-DE. The labeling reactions were performed in the dark on ice for 30 min and then quenched with a 20-fold molar ratio excess of free L-lysine to dye for 10 min. The differentially Cy3- and Cy5-labeled samples were then mixed with the Cy2-labeled internal standard and reduced with dithiothreitol for 10 min. IPG buffer, pH3-10 nonlinear (2% (v/v), GE Healthcare) was added and the final volume was adjusted to 450 μl with 2D-lysis buffer for rehydration. The rehydration process was performed with immobilized nonlinear pH gradient (IPG) strips (pH3-10, 24 cm) which were later rehydrated by CyDye- labeled samples in the dark at room temperature overnight (at least 12 hours). Isoelectric focusing was then performed using a Multiphor II apparatus (GE Healthcare) for a total of 62.5 kV-h at 20ºC. Strips were equilibrated in 6M urea, 30% (v/v) glycerol, 1% SDS (w/v), 100 mM Tris-HCl (pH8.8), 65 mM dithiothreitol for 15 min and then in the same buffer containing 240 mM iodoacetamide for another 15 min. The equilibrated IPG strips were transferred onto 26 x 20-cm 12.5% polyacrylamide gels casted between low fluorescent glass plates. The strips were overlaid with 0.5% (w/v) low melting point agarose in a running buffer containing bromophenol blue. The gels were run in an Ettan Twelve gel tank (GE Healthcare) at 4 Watt per gel at 10ºC until the dye front had completely run off the bottom of the gels. Afterward, the fluorescence 2-DE were scanned directly between the low fluorescent glass plates using an Ettan DIGE Imager (GE Healthcare). This imager is a charge-coupled device-based instrument that enables scanning at different wavelengths for Cy2-, Cy3-, and Cy5-labeled samples. Gel analysis was performed using DeCyder 2-D Differential Analysis Software v7.0 (GE Healthcare) to co-detect, normalize and quantify the protein features in the images. Features detected from non-protein sources (e.g. dust particles and dirty backgrounds) were filtered out. Spots displaying a $\geqq$ 1.3 average-fold increase or decrease in abundance with a p-value < 0.05 were selected for protein identification.

2.4 *Protein staining*

Colloidal coomassie blue G-250 staining was used to visualize CyDye-labeled protein features in 2-DE. Bonded gels were fixed in 30% v/v ethanol, 2% v/v phosphoric acid overnight, washed three times (30 min each) with ddH_2O and then incubated in 34% v/v methanol, 17% w/v ammonium sulphate, 3% v/v phosphoric acid for 1 hr., prior to adding 0.5g/liter

coomassie blue G-250. The gels were then left to stain for 5-7 days. No destaining step was required. The stained gels were then imaged on an ImageScanner III densitometer (GE Healthcare), which processed the gel images as .tif files.

2.5 *In-gel digestion*

Excised post-stained gel pieces were washed three times in 50% acetonitrile, dried in a SpeedVac for 20 min., reduced with 10 mM dithiothreitol in 5 mM ammonium bicarbonate pH 8.0 for 45 min at 50°C and then alkylated with 50 mM iodoacetamide in 5 mM ammonium bicarbonate for 1 hr. at room temperature in the dark. The gel pieces were then washed three times in 50% acetonitrile and vacuum-dried before reswelling with 50 ng of modified trypsin (Promega) in 5 mM ammonium bicarbonate. The pieces were then overlaid with 10 µl of 5 mM ammonium bicarbonate and trypsinized for 16 hr at 37°C. Supernatants were collected, peptides were further extracted twice with 5% trifluoroacetic acid in 50% acetonitrile and the supernatants were pooled. Peptide extracts were vacuum-dried, resuspended in 5 µl ddH_2O, and stored at -20°C prior to MS analysis.

2.6 *Protein identification by MALDI-TOF MS*

Extracted proteins were cleaved with a proteolytic enzyme to generate peptides, then a peptide mass fingerprinting (PMF) database search following MALDI-TOF MS analysis was employed for protein identification. Briefly, 0.5 µl of tryptic digested protein sample was first mixed with 0.5 µl of a matrix solution containing α-cyano-4-hydroxycinammic acid at a concentration of 1 mg in 1 ml of 50% acetonitrile (v/v) / 0.1% trifluoroacetic acid (v/v), spotted onto an anchorchip target plate (Bruker Daltonics) and dried. The peptide mass fingerprints were acquired using an Autoflex III mass spectrometer (Bruker Daltonics) in reflector mode. The algorithm used for spectrum annotation was SNAP (**S**ophisticated **N**umerical **A**nnotation **P**rocedure). This process used the following detailed metrics: Peak detection algorithm: SNAP; Signal to noise threshold: 25; Relative intensity threshold: 0%; Minimum intensity threshold: 0; Maximal number of peaks: 50; Quality factor threshold: 1000; SNAP average composition: Averaging; Baseline subtraction: Median; Flatness: 0.8; MedianLevel: 0.5. The spectrometer was also calibrated with a peptide calibration standard (Bruker Daltonics) and internal calibration was performed using trypsin autolysis peaks at *m/z* 842.51 and *m/z* 2211.10. Peaks in the mass range of *m/z* 800-3000 were used to generate a peptide mass fingerprint that was searched against the Swiss-Prot/TrEMBL database (release on 05-Oct-10) with 521016 entries using Mascot software v2.3.02 (Matrix Science, London, UK). The following parameters were used for the search: *Homo sapiens*; tryptic digest with a maximum of 1 missed cleavage; carbamidomethylation of cysteine, partial protein N-terminal acetylation, partial methionine oxidation and partial modification of glutamine to pyroglutamate and a mass tolerance of 50 ppm. Identification was accepted based on significant MASCOT Mowse scores ($p<0.05$), spectrum annotation and observed versus expected molecular weight and p*I* on 2-DE.

3. Results and Discussion

3.1 *Plasma sample preparation*

Because albumin and immunoglobulin G account for around 70-80% proteins in human plasma, these high abundant proteins are the obstacle for 2-DE analysis. Accordingly, removing the high-abundance albumin and immunoglobulin G from plasma samples can increase the viability of the low- and middle-abundance proteins and enable for accurate analysis. In this study, the high abundance proteins were removed before performing 2D-DIGE analysis, and the effect of depletion was evaluated by 1D-SDS-PAGE (data not shown).

3.2 *2D-DIGE and mass spectrometry analysis of the immunodepleted plasma proteome*

In order to study the alteration of plasma protein in the type 2 diabetic patients, comparative proteomics analysis was performed between type 2 diabetic patients and healthy donors. The 2-DE images of the samples of two groups were minimally labeled with Cy3 and Cy5 dyes and distributed to each gel. A pool of both samples was also prepared for labeling with Cy2 as an internal standard to run on all gels to facilitate image matching across gels. The plasma samples arrangement for a triplicate 2D-DIGE experiment is shown in Figure 1A. Thus, the triplicate samples resolved in different gels can be quantitatively analyzed by means of the internal standard on multiple 2-DE. After resolving protein samples with 2D-DIGE technique, the DeCyder image analysis software indicated that 65 protein features were showing greater than 1.3-fold change in expression level with student t-test (*p*-value) less than 0.05. MALDI-TOF MS identification revealed that 34 proteins were differentially expressed (Figure 1B, Figure 2 and Supplementary Table 1). With the basis of a Swiss-Prot search and KEGG pathway analysis, numerous potential biological functions and subcellular locations of the identified proteins were determined. Most of these identified proteins belong to cytoplasm proteins (29%), nuclear proteins (26%) and mitochondrial proteins (12%), and these proteins are found to be involved in metabolism (20%) and growth regulation (17%) (Figure 3).

Representative examples for the evaluation by DeCyder of alteration in spot intensities using the 2D-DIGE system are displayed in Figure 4. To display visually alterations in corresponding spot intensity proportions, selected identified spots are shown as 3-D images as well as the associated graph views of standardized abundances of the selected spots (Figure 4).

4. Concluding remarks

Proteomic analysis of the human diseases usually adopt a comparative method that is defined by the differential expression of the proteins under different disease states. Our 2D-DIGE / MALDI-TOF analysis revealed 34 altered expression

A

	Cy2 (50ug)	Cy3 (50ug)	Cy5 (50ug)
Gel1	pool	Normal	DM2
Gel2	pool	DM2	Normal
Gel3	pool	Normal	DM2

B

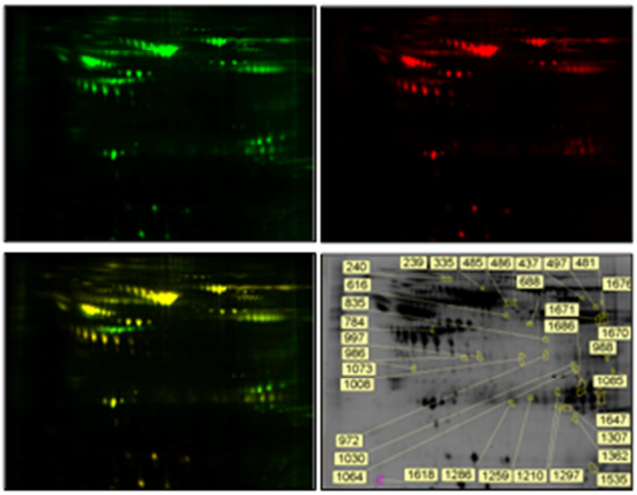


Figure 1. 2D-DIGE analysis of type 2 diabetes mellitus-induced differentially expressed proteins. *(A)* Plasma samples arrangement for a triplicate 2D-DIGE experiment. *(B)* Plasma samples (50 mg each) were labeled with Cy-dyes and separated using 24 cm, pH 3-10 non-linear IPG strips. 2D-DIGE images of the plasma samples from type 2 diabetic patients and healthy individuals at appropriate excitation and emission wavelengths were shown (Upper images) as well as overlaid pseudo-colored image processed with ImageQuant Tool (GE Healthcare) (left bottom). The differentially expressed identified protein features are annotated with spot numbers (right bottom).

of plasma proteins corresponding to 31 unique plasma proteins (Supplementary Table 1). A majority of altered proteins belong to three major functional groups, metabolism, growth regulation and transportation, while other affected categories included cytoskeleton, immune regulation and membrane trafficking (Figure 3). Of these, 20% participate in various metabolic pathways including 5'-AMP-activated protein kinase subunit beta-1 (regulation of energy production and consumption), 6-phosphofructokinase (glycolysis), carnitine O-palmitoyltransferase I (long-chain fatty acid transportation), choline/ethanolamine kinase (biosynthesis of phosphatidylcholine / phosphatidylethanolamine), steroid 21-hydroxylase (biosynthesis of the steroid hormones aldosterone and cortisol), glutamine-dependent NAD(+) synthetase (biosynthesis of NAD). 5'-AMP-activated protein kinase subunit beta-1 is one of the main subunits of AMPK proteins which plays as a key player in regulating energy metabolism and renders it at the center role in studies of the metabolic diseases including diabetes. AMPK is also reported as a critical molecule to maintain glucose homeostasis and its activation and expression is a positive regulator to reduce the concentration of blood glucose and facilitate the oxidation of fatty acids and glycolysis [13]. In current study, we have observed the down-regulated of AMPK-beta subunit, carnitine O-palmitoyltransferase I and 6-phosphofructokinase in type 2 diabetes mellitus plasma suggesting the up-stream regulators might inhibit AMPK expression subsequently by hampered the transportation of fatty acid into mitochondria for beta-oxidation via the inhibition of carnitine O-palmitoyltransferase I and down-regulated glycolysis via the inhibition of 6-phosphofructokinase in this disease.

Apolipoprotein A-I is a component of the high density lipoprotein responsible for the cholesterol transport into the liver [14]. Apolipoprotein A-I is also associated with diabetes and demonstrates a downregulation of apolipoprotein A-I in diabetic patients [15-18]. Our current 2D-DIGE analysis is consistent with the previous observations. Ficolins are a group of proteins containing collagen-like domains and fibrinogen-like domains. In addition, ficolins have a calcium-independent lectin activity and are able to activate the complement pathway to improve host defense activity [19]. In our current work and recent plasma proteomic studies demonstrated that ficolin-3 was significantly correlated with type 2 diabetes mellitus and supposed to play an important role in the chronic low-grade inflammation and stimulation of the innate immune system associated with type 2 diabetes mellitus [16,20]. Other identified proteins including LIM homeobox protein, 6-phosphofructokinase, steroid 21-hydroxylase and Rho GTPase-activating protein 25 have been well-studied in type 2 diabetes mellitus [21-23] suggesting the current 2D-DIGE playing a powerful complementary role in the assumed biomarker discovery and disease studies. On the other hand,

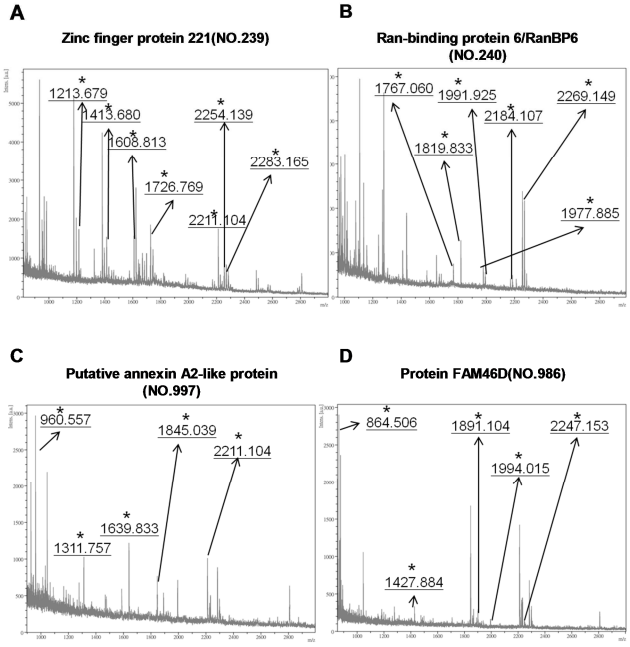


Figure 2. Peptide mass fingerprint of identified proteins *(A)* Zinc finger protein 221; *(B)* Ran-binding protein 6 / RanBP6; *(C)* Putative annexin A2-like protein; *(D)* Protein FAM46D. Peptides contributes to protein identifications are marked with asterisk (*).

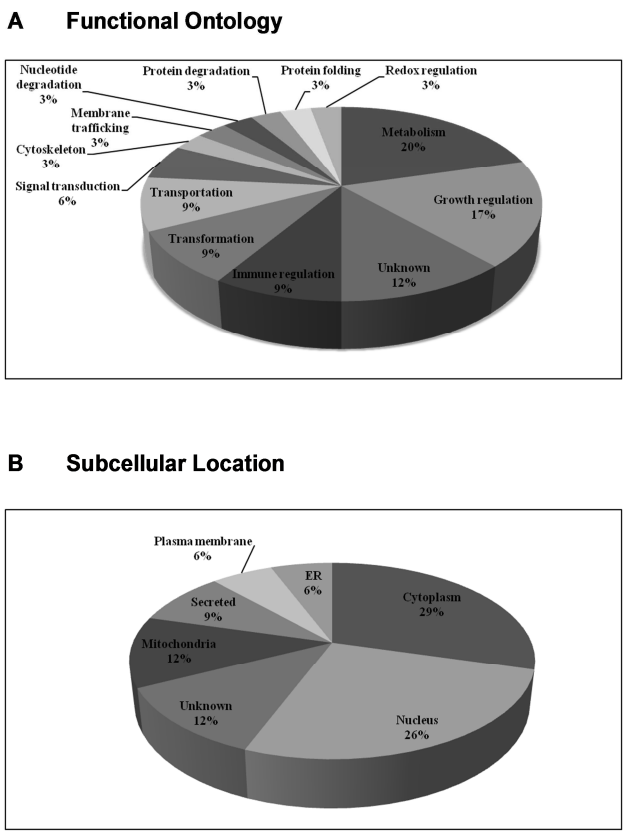


Figure 3. Percentage of plasma proteins identified from albumin and imunnoglobulin G depleted plasma by 2D-DIGE / MALDI-TOF MS for type 2 diabetes mellitus according to their biological functions *(A)* and subcellular location *(B)*.

our current research also identifies alpha-(1,3)-fucosyltransferase, SCaMC-3, calpain-7, choline/ ethanolamine kinase, paraneoplastic antigen-like protein 6B, ran-binding protein 6/RanBP6, Sulfide:quinone oxidoreductase glutamine-dependent NAD(+) synthetase as putative plasma marker proteins of the type 2 diabetes mellitus. Further investigations would be necessary to validate their expression in plasma and elucidate their mechanistic role in type 2 diabetes mellitus.

The isolation of low-abundance proteins from plasma is frequently complicated because of the existed of high abundance proteins such as serum albumin and immunoglobulin. Serum albumin and immunoglobulin is the most (50-70%) and second (10-20%) abundant protein in plasma, respectively. These proteins will mask the low-abundant proteins and limit the amount of the total plasma proteins that can be applied and resolved by proteomic analysis. To removal of these two high-abundant proteins, we used albumin and IgG depletion kit from Sigma-Aldrich, which contains prepacked spin columns allows removal of the most of both serum albumin and immunoglobulin. At the same time, we used trichloroacetic acid / acetone to precipitate, desalt and enrich the plasma proteins to make them well resolved by 2D-DIGE. Our results demonstrated that this strategy did make the plasma proteins well prepared and separated.

In conclusion, the quantitative plasma proteomics analysis provided a valuable impact for type 2 diabetes mellitus research. Our quantitative proteomic approach has identified numerous previous reported plasma markers of type 2 diabetes mellitus such as apolipoprotein A-I and ficolin-3. Additionally, we have presented several putative type 2 diabetes mellitus biomarkers which may be associated with the progression and development of the disease and has a potential to serve as a useful tool for monitoring the course of the disease. The potential of utilizing these markers for screening and treating type 2 diabetes mellitus warrants further investigation.

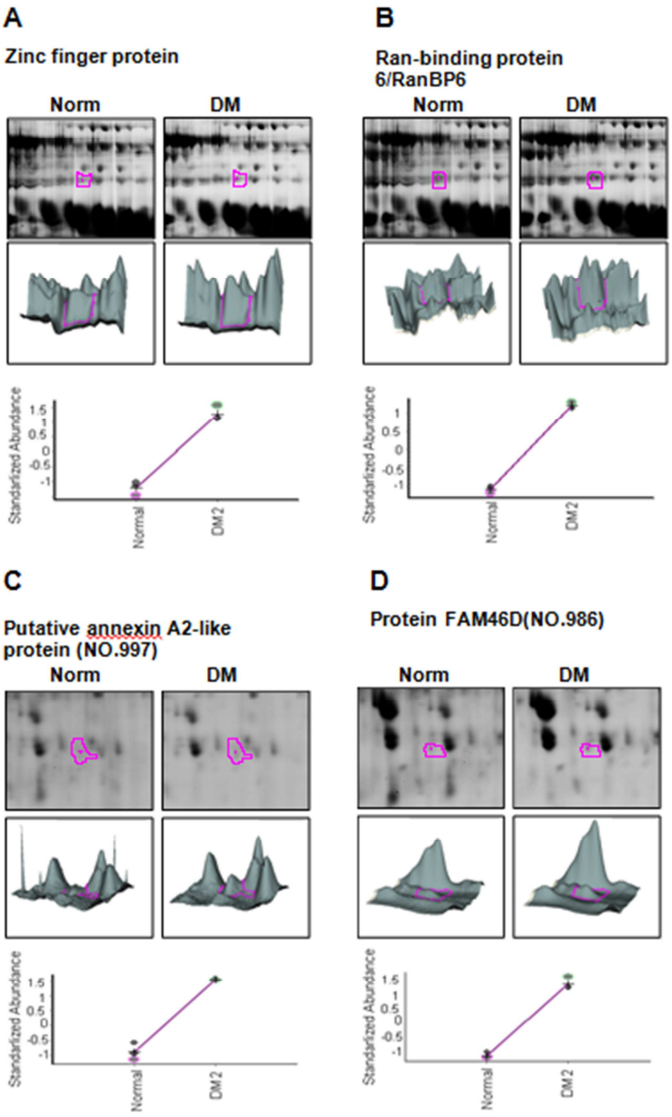


Figure 4. Representative images of the identified spots ((A) Zinc finger protein 221; (B) Ran-binding protein 6 / RanBP6; (C) Putative annexin A2-like protein; (D) Protein FAM46D) displaying type 2 diabetes mellitus–dependent protein expression changes. The levels of these proteins were visualized by 2-DE images (top panels), three-dimensional spot images (middle panels) and protein expression map (bottom panels).

5. Supplementary material

Supplementary material regarding this manuscript is online available in the web page of JIOMICS.

http://www.jiomics.com/index.php/jio/rt/suppFiles/44/0

Acknowledgements

This work was supported by grant (NSC 99-2311-B-007-002) from the National Science Council, Taiwan, NTHU Booster grant (99N2908E1) from the National Tsing Hua University and grant (VGHUST99-P5-22) Veteran General Hospitals University System of Taiwan. The authors also thank the grant supported from the Chiayi Christian Hospital, Chiayi, Taiwan.

References

1. Korc, M., Mol. Cell Proteomics. 2003, 2, 399-404.
2. Horikawa, Y., Yamasaki, T., Nakajima, H., Shingu, R. et al., Horm. Metab Res. 2003, 35, 308-312.
3. Vendrell, J., Fernandez-Real, J. M., Gutierrez, C., Zamora, A. et al., Atherosclerosis 2003, 167, 257-264.
4. On, Y. K., Park, H. K., Hyon, M. S., Jeon, E. S., Circ. J. 2007, 71, 868-873.
5. Timms, J. F., Cramer, R., Proteomics. 2008, 8, 4886-4897.
6. Westermeier, R., Scheibe, B., Methods Mol. Biol. 2008, 424, 73-85.
7. Marouga, R., David, S., Hawkins, E., Anal. Bioanal. Chem. 2005, 382, 669-678.
8. Lai, T. C., Chou, H. C., Chen, Y. W., Lee, T. R. et al., J. Proteome. Res. 2010, 9, 1302-1322.
9. Chou, H. C., Chen, Y. W., Lee, T. R., Wu, F. S. et al., Free Radic. Biol. Med. 2010, 49, 96-108.
10. Huang, H. L., Hsing, H. W., Lai, T. C., Chen, Y. W. et al., J. Biomed. Sci. 2010, 17, 36.
11. Chan, H. L., Gharbi, S., Gaffney, P. R., Cramer, R. et al., Proteomics. 2005, 5, 2908-2926.
12. Gharbi, S., Gaffney, P., Yang, A., Zvelebil, M. J. et al., Mol Cell Proteomics 2002, 1, 91-8.
13. Zhang, B. B., Zhou, G., Li, C., Cell Metab 2009, 9, 407-416.
14. Rader, D. J., Am. J. Cardiol. 2003, 92, 42J-49J.
15. Rohrer, L., Hersberger, M., von Eckardstein, A., Curr. Opin. Lipidol. 2004, 15, 269-278.
16. Li, R. X., Chen, H. B., Tu, K., Zhao, S. L. et al., PLoS. One. 2008, 3, e3224.
17. Liu, X., Feng, Q., Chen, Y., Zuo, J. et al., J. Proteome. Res. 2009, 8, 1255-1262.
18. Rao, A. A., Sridhar, G. R., Das, U. N., Med. Hypotheses 2007, 69, 1272-1276.
19. Matsushita, M., Kuraya, M., Hamasaki, N., Tsujimura, M. et al., J. Immunol. 2002, 168, 3502-3506.
20. Kim, H. J., Cho, E. H., Yoo, J. H., Kim, P. K. et al., J. Proteome. Res. 2007, 6, 735-743.
21. Ahmed, M., Neville, M. J., Edelmann, M. J., Kessler, B. M., Karpe, F., Obesity. (Silver. Spring) 2010, 18, 27-34.
22. Matsumoto, T., Kobayashi, T., Ishida, K., Taguchi, K., Kamata, K., Br. J. Pharmacol. 2010, 160, 1092-1104.
23. Peterson, P., Salmi, H., Hyoty, H., Miettinen, A. et al., Clin. Immunol. Immunopathol. 1997, 82, 37-42.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.46

A novel phosphate-affinity bead with immobilized Phos-tag for separation and enrichment of phosphopeptides and phosphoproteins

Emiko Kinoshita-Kikuta[1], Atsushi Yamada[1], Chika Inoue[1], Eiji Kinoshita*[1], Tohru Koike[1].

[1]Department of Functional Molecular Science, Graduate School of Biomedical Sciences, Hiroshima University, Hiroshima, Japan.

Received: 29 October 2010 **Accepted:** 1 December 2010 **Available Online**: 2 December 2010

ABSTRACT

A simple and efficient method was developed for separating and enriching phosphoproteins from crude biological samples containing solubilized cellular proteins by immobilized zinc(II) affinity chromatography. The phosphate-binding site of the affinity gel is an alkoxide-bridged dinuclear zinc(II) complex, Phos-tag, which is linked to a hydrophilic vinylic polymer bead. A novel phosphate-affinity bead (Phos-tag Toyopearl) was prepared by reaction of N-hydroxysuccinimide-activated Toyopearl AF-Carboxy-650M gel with a 2-aminoethylcarbamoyl derivative of Phos-tag. Phosphopeptides were retrieved quantitatively and selectively on a μL-scale column at room temperature. The column was stable for long-term storage and could be reused many times. The technique was used to separate and enrich phosphoproteins from an epidermal growth factor-stimulated human epidermoid carcinoma A431 cell lysate. The operations necessary for 1-mL-scale open-column chromatography were conducted at a physiological pH during 1 h. The strong enrichment of the phosphoproteins into the eluted fraction was evaluated by gel electrophoresis, followed by Western blotting with Phos-tag Biotin and several antibodies, Pro-Q Diamond phosphoprotein gel staining, and mass spectrometry.

Keywords: Affinity chromatography; Phosphopeptide; Phosphoprotein; Phosphoproteomics; Phosphorylation; Phos-tag.

Abbreviations

2-DE, two-dimensional gel electrophoresis; **CBB**, Coomassie Brilliant Blue; **ECL**, enhanced chemiluminescence; **EGF**, epidermal growth factor; **HPLC**, high-performance liquid chromatography; **HRP**, horseradish peroxidase; **IEF**, isoelectric focusing; **IgG**, immunoglobulin G; **IMAC**, immobilized metal ion affinity chromatography; **MALDI-TOF**, matrix-assisted laser desorption/ionization-time of flight; **MES**, 2-morpholin-4-ylethanesulfonic acid; **MOAC**, metal oxide affinity chromatography; **MS**, mass spectrometry; **NHS**, N-hydroxysuccinimide; **PAGE**, polyacrylamide gel electrophoresis; pI, isoelectric point; **PVDF**, poly(vinylidene difluoride); **RIPA**, radio-immunoprecipitation assay; **SDS**, sodium dodecyl sulfate; **TFA**, trifluoroacetic acid.

1. Introduction

Phosphorylation of proteins dramatically increases the diversity of genetically encoded proteins and plays a key role in regulating the function, localization, binding specificity, and stability of target proteins [1]. To evaluate the role of phosphorylation, it is necessary to identify phosphorylated amino acid residues, to analyze activities of kinases and phosphatases, and to elucidate interactions among proteins. Therefore, techniques for the specific separation and enrichment of native phosphoproteins have attracted considerable interest in relation to studies on the phosphoproteome in the fields of biology and medicine.

A number of studies have been reported on the enrichment of phosphoproteins from biological samples such as cell lysates. The method that is most frequently used is affinity chromatography using immobilized antibodies to phosphorylated amino acids [2–4]. However, the antibody-based procedure has a drawback in that it is incapable of comprehensive enrichment of phosphoproteins. Other widely accepted

***Corresponding author:** Eiji Kinoshita, Postal address: Kasumi 1-2-3, Hiroshima 734-8553, Japan, Tel.: 81-82-257-5281; Fax: 81-82-257-5336 ;Email Address: kinoeiji@hiroshima-u.ac.jp;

methods are immobilized metal ion affinity chromatography (IMAC) [5] and metal oxide affinity chromatography (MOAC) [6], which provide more-comprehensive enrichment. However, these techniques have a problem in that the solubility of proteins is reduced under the acidic conditions that are required in the workup for chromatography. These procedures are, however, generally useful for enriching phosphopeptides produced by digestion of phosphoproteins. Recently, improvements in the specificity of IMAC/MOAC have been accomplished in a number of ways for mass spectrometry (MS)-based studies on the phosphoproteome [7–13].

In 2005, we reported a novel type of immobilized zinc(II) affinity chromatography, known as Zn(II)-IMAC, in which a phosphate-affinity gel, Phos-tag Agarose, is used for the separation and enrichment of phosphoproteins and phosphopeptides [14]. The gel consists of biopolymer beads (cross-linked agarose) containing an immobilized alkoxide-bridged dinuclear zinc(II) complex {Phos-tag: 1,3-bis[bis(pyridin-2-ylmethyl)amino]propan-2-olato dizinc(II) complex}. The Phos-tag molecule, which was synthesized to mimic the active center of an alkaline phosphatase, acts as a phosphate-binding tag molecule in aqueous solutions at neutral pH {e.g., K_d = 25 nM for the phenyl phosphate dianion [PhOP(=O)(O^-)$_2$]} [15]. The anion selectivity indexes of the phenyl phosphate dianion against SO_4^{2-}, CH_3COO^- (AcO^-), Cl^-, and the diphenyl phosphate monoanion [$(PhO)_2P(=O)O^-$] at 25 °C are 5.2×10^3, 1.6×10^4, 8.0×10^5, and $>2 \times 10^6$, respectively. The characteristic of Phos-tag molecule permits a comprehensive enrichment of various biological phosphorylated compounds, providing a major advantage in obtaining information on intact native phosphoproteins present in complex samples of cellular lysates [16,17]. However, the biopolymer beads are physically and chemically unstable, and they are therefore unsuitable for multiple use or long-term storage.

To increase the opportunities for analyses of phosphoproteins in a range of applications (from, for example, processing of microvolume samples by column chromatography to high-speed processing of high-volume samples in fast-flow rate liquid chromatography) [18,19], we have developed a more-stable phosphate-affinity bead by using an amino-pendant Phos-tag molecule and a hydrophilic vinylic polymer gel (Toyopearl; Tosoh, Tokyo, Japan) [20–22]. We have also demonstrated the efficiency and advantages of our new method by using the phosphate-affinity beads (Phos-tag Toyopearl) to analyze phosphopeptides and phosphoproteins in model experiments.

2. Material and methods

2.1 *Reagents*

Phos-tag Agarose [14,16,17] and Phos-tag Biotin [23–25] are commercially available from Wako Pure Chemical Industries, Ltd. (Osaka, Japan). Toyopearl AF-Carboxy-650M, a proprietary vinylic resin functionalized with pendant carboxyl groups, was purchased from Tosoh (Tokyo, Japan). Bovine β-casein, chicken egg white ovalbumin, epidermal growth factor (EGF), proteomics-grade trypsin, and PHOS-select Iron Affinity Gel were purchased from Sigma-Aldrich (St. Louis, MO, USA). Src peptide 521–533 and phosphorylated Src peptide 521–533 were purchased from CalBiochem (La Jolla, CA, USA). Phosphorylated PKC substrate, 4A/4B peptide, insulin receptor 1142–1153, kinase domain of insulin receptor-2, and kinase domain of insulin receptor-5 were purchased from AnaSpec (San Jose, CA, USA). PKC substrate peptide was purchased from Stressgen (San Diego, CA, USA). [Ala-97]-RII 81–99 (PKA substrate), RII phosphopeptide (calcineurin substrate), MAP 177–189 pT/pY peptide, and EGFR 988–998 (PTP1B substrate) were purchased from Enzo Life Sciences (Plymouth Meeting, PA, USA). The ECL Advance Western Blotting Detection Kit, horseradish peroxidase (HRP)-conjugated anti-mouse immunoglobulin G (IgG) antibody, HRP-conjugated anti-rabbit IgG antibody, and HRP-conjugated streptavidin were purchased from GE Healthcare Bio-Sciences (Piscataway, NJ, USA). Pro-Q Diamond phosphoprotein gel stain, SYPRO Ruby protein gel stain, and anti-pSer antibody were purchased from Invitrogen (Carlsbad, CA, USA). Anti-pMAPK1/2 antibody (clone 12D4; against phosphorylated Thr202/Tyr204), anti-MAPK1/2 antibody, anti-pShc antibody (against phosphorylated Tyr317), and anti-pErbB-2/HER-2 antibody (against phosphorylated Tyr1248) were purchased from Millipore (Billerica, MA, USA). Anti-pMAPK substrates (PXTP) antibody (clone 46G11) was purchased from Cell Signaling Technology (Danvers, MA, USA). A PhosphoProtein Purification Kit was purchased from Qiagen (Hilden, Germany). Titansphere TiO was purchased from GL Sciences (Tokyo, Japan). A protein assay kit and the ReadyPrep 2-D Clean Up Kit were purchased from Bio-Rad Laboratories (Hercules, CA, USA). An EzApply 2D Kit and the first dimensional precast agarGEL (A-M310, pH 3–10) were purchased from Atto (Tokyo, Japan).

2.2 *Preparation of the immobilized Zn(II) Phos-tag Toyopearl affinity beads*

Phos-tag Toyopearl without any bound zinc(II) was prepared by a condensation reaction between an amino-pendant Phos-tag ligand (75 μmol, Fig. 1) and Toyopearl AF-Carboxy-650M (5.0 mL, ~100 μmol carboxyl group per mL of gel) by a similar method to that reported previously [14]. In the presence of 0.50 mmol N-hydroxysuccinimide (NHS; 1-hydroxypyrrolidine-2,5-dione) and 0.50 mmol N-[3-(dimethylamino)propyl]-N'-ethylcarbodiimide hydrochloride, the amide formation proceeded quantitatively, as evaluated from an analysis of the amount of unbound Phos-tag molecule by high-performance liquid chromatography (HPLC) [eluent: 90% (v/v) aqueous HClO4 (pH 2) and 10% (v/v) CH3CN] at 254 nm. The unreacted NHS ester groups on the beads were hydrolyzed overnight in 0.10 M NaHCO3–Na2CO3 (pH 10.5, 10 mL) to reform carboxylate groups. The beads were then washed five times with aqueous 50% (v/v) EtOH (10 mL) and five times with 10 mM 2-morpholin-4-

ylethanesulfonic acid (MES)–NaOH (pH 6.0, 10 mL). Finally, the beads were incubated in a buffer solution of 10 mM MES–NaOH (pH 6.0, 10 mL) containing 0.10 M Zn(OAc)2 for 1 h at room temperature to form the zinc(II)-bound beads. The resulting zinc(II)-bound beads were washed five times with 0.10 M Tris–AcOH (pH 7.4, 10 mL) containing 20% (v/v) propan-2-ol, and the washed Phos-tag Toyopearl beads (15 μmol/mL-gel of Phos-tag) were stored as a 50% (v/v) slurry in the same buffer at 4 °C (see Figure 1).

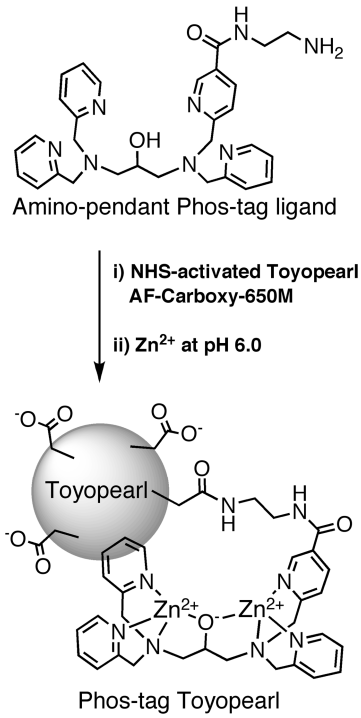


Figure 1. Preparation of Phos-tag Toyopearl, the novel Zn(II)-IMAC bead, by using a Phos-tag derivative having a 2-aminoethylcarbamoyl group (amino-pendant Phos-tag) and NHS-activated Toyopearl AF-Carboxy-650M gel.

2.3 *Preparation of the lysed proteins from the cultured cells*

A431 human epidermoid carcinoma cell line was obtained from the Cell Resource Center for Biomedical Research, Institute of Development, Aging, and Cancer at Tohoku University (Japan). The cells were grown in an RPMI1640 medium containing 10% (v/v) fetal bovine serum, 100 units/mL penicillin, and 100 μg/mL streptomycin under a humidified atmosphere of 5% CO_2 and 95% air at 37 °C. The cells (7×10^6) were placed in the same medium on a 100-mm culture plate. When the cells had adhered to the plate, the medium was removed and replaced with serum-free medium. After incubation for 16 h, the cells were stimulated with 250 ng/mL of EGF for 5 min. To terminate the stimulation, the medium was removed and the remaining cells were rinsed with Tris-buffered saline at room temperature. The saline was then removed and the culture plate was placed on ice. The cells were exposed to 0.30 mL of a cold radioimmunoprecipitation assay (RIPA) buffer consisting of 50 mM Tris–HCl (pH 7.4), 0.15 M NaCl, 0.25% (w/v) sodium deoxycholate, 1.0% (v/v) Nonidet P-40, 1.0 mM EDTA, 1.0 mM phenylmethanesulfonyl fluoride, 1.0 μg/mL aprotinin, 1.0 μg/mL leupeptin, 1.0 μg/mL pepstatin, 1.0 mM Na_3VO_4, and 1.0 mM NaF. The plate was gently rocked for 15 min on ice, and the adherent cells were then removed from the plate with a cell scraper. The resulting suspension was transferred to a microcentrifuge tube. The plate was washed with 0.20 mL of an RIPA buffer, and the washing solution was combined with the first suspension in the microcentrifuge tube. The mixed sample was incubated for 60 min on ice and then centrifuged at 10,000 *g* for 10 min at 4 °C. The supernatant fluid was used as the cell lysate. The concentration of solubilized proteins was adjusted to 2.0 mg/mL by dilution with an appropriate amount of the RIPA buffer. Quantification of proteins was performed by using a Bio-Rad protein assay kit and a Biotrak II microplate reader (GE Healthcare Bio-Sciences). The lysed protein samples were subjected to phosphate-affinity chromatography with a spin column and with an open column as described below in Sections 2.4 and 2.12.

2.4 *Phosphate-affinity chromatography with a spin column*

Phos-tag Toyopearl beads (a 50-μL bed of compressed gel) were placed in a sample reservoir (an Ultrafree-MC centrifugal filter unit; Millipore). The filter unit was centrifuged at 2000 *g* for 20 s to remove the storage buffer [0.10 M Tris–acetic acid (AcOH) (pH 7.5) containing 20% (v/v) propan-2-ol], and the supernatant was discarded. To form the zinc(II)-bound Phos-tag Toyopearl beads, a balancing buffer (0.10 mL) was placed in the sample reservoir. For separation of protein samples, the balancing buffer consisted of 0.10 M Tris–AcOH (pH 7.5), 0.50 M NaOAc, and 10 μM $Zn(OAc)_2$, whereas for the separation of peptide samples, it consisted of 10 Mm MES–NaOH (pH 6.0), 0.10 M NaCl, 1.0 mM disodium oxalate, and 10 μM $Zn(OAc)_2$. The filter unit was centrifuged at 2000 *g* for 20 s, and the supernatant buffer was discarded. A binding/washing buffer (0.10 mL) was placed in the sample reservoir. For the separation of protein samples, the binding/washing buffer consisted of 0.10 M Tris–AcOH (pH 7.5) and 0.50 M NaOAc, whereas for the separation of peptide samples it consisted of 10 mM MES–NaOH (pH 6.0), 0.10 M NaCl, and 1.0 mM disodium oxalate. The unit was then centrifuged at 2000 *g* for 20 s and the supernatant buffer was discarded; this washing operation was repeated twice.

For the preparation of the sample of proteins, the EGF-stimulated A431 cell lysate (20 μg proteins in 10 μL of an RIPA buffer) was diluted with 40 μL of the binding/washing buffer for separation of phosphoproteins. For the preparation of the sample of peptides, commercially available peptides (5.0 nmol phosphopeptides and 5.0 nmol nonphosphopeptides) were dissolved in the binding/washing buffer for separation of phosphopeptides (0.10 mL). Each sample solution was added to the sample reservoir and allowed to incubate for 5 min. Next, the filter unit was centrifuged at 2000 *g* for 20 s, and the filtrate was collected as a flow-through fraction. Binding/washing buffer (0.10 mL) was then added to the sample reservoir to wash the Phos-tag Toyopearl beads. The

filter unit was again centrifuged at 2000 *g* for 20 s, and the filtrate was collected as a washing fraction containing unbound proteins/peptides. For the separation of the protein sample, this washing operation was repeated twice, and the filtrate was collected as washing fractions, whereas for separation of the peptide sample, the washing operation was repeated three times, and the filtrate was collected as washing fractions. The flow-through and washing fractions were combined and used for subsequent analysis.

To elute the gel-bound proteins or peptides, an elution buffer (0.10 mL) was added to the sample reservoir. For proteins, the elution buffer consisted of 0.10 M Tris–AcOH (pH 7.5), 1.0 M NaCl, and 10 mM NaH_2PO_4–NaOH (pH 7.5), whereas for peptides it consisted of 0.10 M NaH_2PO_4–NaOH (pH 7.5). The filter unit was centrifuged at 2000 *g* for 20 s, and the filtrate was collected as an elution fraction. For the separation of the protein sample, the eluting operation was repeated twice, and the filtrate (0.20 mL) was collected as two elution fractions, whereas for the separation of the peptide sample, the eluting operation was repeated four times. In the case of separation of the protein sample, to analyze the proteins remaining in the column, a column-washing buffer (0.10 mL) consisting of 0.10 M Tris–AcOH (pH 7.5), 1.0% (w/v) sodium dodecyl sulfate (SDS), and 10 mM EDTA was added to the sample reservoir. The unit was heated for 5 min at 95 °C and then centrifuged at 2000 *g* for 20 s. The filtrate was collected as a column-washing fraction that contained the residual proteins from the column. Each fraction was desalted and condensed by using a Microcon YM-10 centrifugal filter unit (Millipore) and then dissolved in 20 μL of distilled water. The resulting solution was divided into two fractions of 10 μL each. One of these fractions was analyzed by SDS-polyacrylamide gel electrophoresis (PAGE) with SYPRO Ruby gel staining and the other was analyzed by SDS-PAGE with Western blotting. In the case of the separation of the peptide sample, all the fractions were analyzed by reversed-phase HPLC to estimate the relative recovery of the peptides. All the spin column chromatography operations were performed at room temperature.

2.5 *Reversed-phase HPLC*

HPLC for analysis of peptides was performed by using a Tosoh LC-8020 Model II multi-station with a CO-8020 column oven (at 40 °C), a UV-8020 UV detector (at 215 nm), two DP-8020 pumps (operating at a flow speed of 1.0 mL/min), an AS-8021 auto sampler, and a Shiseido reversed-phase column (CAPCELL PAK C18 UG80 S-5, 4.6 mm I.D. × 150 mm) (Tokyo, Japan). The HPLC analysis was conducted by gradient elution with an aqueous solution [from 95% to 60% (v/v)] of 0.10% (v/v) trifluoroacetic acid (TFA) and a CH_3CN solution [from 5% to 40% (v/v)] of 0.10% (v/v) TFA for 30 min. The HPLC retention times of each peptide are as follows: phosphorylated PKC substrate, 6.7 min; kinase domain of insulin receptor-5, 9.4 min; phosphorylated Src peptide 521–533, 13.4 min; kinase domain of insulin receptor-2, 14.5 min; PKC substrate peptide, 15.6 min; insulin receptor 1142–1153, 17.1 min; Src peptide 521–533, 17.5 min; EGFR 988–998, 18.1 min; MAP 177–189 pT/pY peptide, 21.7 min; RII phosphopeptide, 26.7 min; 4A/4B peptide, 27.1 min; and [Ala-97]-RII 81–99, 27.7 min.

2.6 *SDS-PAGE*

SDS-PAGE, conducted according to the Laemmli's method [26], was usually performed at 35 mA/gel and room temperature in a 1-mm-thick, 9-cm-wide, and 9-cm-long gel on a PAGE apparatus (AE6500; Atto). The gel consisted of a stacking gel [4.0% (w/v) polyacrylamide, 125 mM Tris–HCl (pH 6.8), and 0.10% (w/v) SDS] and a separating gel [7.5–12.5% (w/v) polyacrylamide, 375 mM Tris–HCl (pH 8.8), and 0.10% (w/v) SDS]. The acrylamide stock solution was prepared as a 29:1 (w/w) mixture of acrylamide and *N*,*N*'-methylenebisacrylamide. The electrophoresis running buffer (pH 8.4) consisted of 25 mM Tris and 192 mM glycine containing 0.10% (w/v) SDS. Each sample was prepared by mixing the separated fraction with half its volume of an SDS-PAGE loading buffer consisting of 195 mM Tris–HCl (pH 6.8), 9.0% (w/v) SDS, 15% (v/v) 2-mercaptoethanol, 30% (v/v) glycerol, and 0.10% (w/v) bromophenol blue [4,4'-(1,1-dioxido-3*H*-2,1-benzoxathiole-3,3-diyl)bis(2,6-dibromophenol)]. All the samples were heated for 5 min at 95 °C before they were loaded onto the gel.

2.7 *Western blotting*

The SDS-PAGE gel was electrotransferred to a poly(vinylidene difluoride) (PVDF) membrane (Fluorotrans W; Nihon Pall, Tokyo, Japan) for 1 h at 100 V by using a Mini Trans-Blot Electrophoretic Transfer Cell (Bio-Rad Laboratories) as a blotting system. The gel was soaked in a solution containing 25 mM Tris, 192 mM glycine, and 10% (v/v) MeOH for 10 min. The protein-blotted PVDF membrane was steeped in a solution containing 10 mM Tris–HCl (pH 7.5), 0.10 M NaCl, and 0.10% (v/v) Tween 20 (TBS-T solution) for 1 h, and then blocked with a solution of 1.0% (w/v) bovine serum albumin in TBS-T for 1 h. The membrane was probed separately with anti-MAPK1/2 antibody, anti-pMAPK1/2 antibody, anti-pSer antibody, anti-pMAPK substrates (PXTP) antibody, anti-pShc antibody, and anti-pErbB-2/HER-2 antibody. The membrane was then probed with HRP-conjugated anti-mouse IgG antibody for detection of the anti-pMAPK antibody, and with HRP-conjugated anti-rabbit IgG antibody for detection of the other primary antibodies. The enhanced chemiluminescence (ECL) signals were detected by means of an ECL Advance Western Blotting Detection Kit and were observed by using an LAS 3000 image analyzer (Fujifilm, Tokyo, Japan). For comprehensive detection of phosphoprotein, we prepared a complex of Phos-tag Biotin and HRP-conjugated streptavidin as described previously [24]. The membrane was probed with the complex for 30 min without blocking, and then washed twice with TBS-T at room temperature for 5 min each time. The ECL signal was detected as described above.

2.8 *Tryptic digestion of β-casein*

Bovine β-casein was subjected to tryptic digestion in a mixed solution of 1.0 mg β-casein in distilled water (0.10 mL) and 0.10 M NH_4HCO_3 (pH 8.1, 0.40 mL) containing 20 μg of proteomics-grade trypsin for 12 h at 37 °C. The trypsin and other large molecules (>10,000 Da) were removed by ultrafiltration through a Microcon YM-10 ultrafilter unit at 14,000 *g* for 40 min. An aqueous solution (0.20 mL) of 0.10% (v/v) TFA was then added to the ultrafiltration unit, and ultrafiltration was performed again as described above. The resulting filtrates were combined (~1.0 mL) and a 10-μL aliquot of the filtrate containing ~10 μg β-casein digest was freeze-dried and stored below 0 °C. The tryptic digestion samples were subjected to phosphate-affinity chromatography on a pipette microtip column as described below (Section 2.10).

2.9 *In-gel tryptic digestion of ovalbumin*

The SDS-PAGE gel bands for ovalbumin (total ~10 μg) were excised with a razor blade and cut into approximately 1–2 mm^3 pieces. The pieces were placed in a microcentrifuge and destained by washing with 0.10 M NH_4HCO_3 (pH 8.1) containing 50% (v/v) CH_3CN (0.50 mL) for 30 min. The waste was discarded and the washing step was repeated three times. The sample proteins were reduced with 0.10 M NH_4HCO_3 (pH 8.1) containing 10 mM EDTA and 10 mM dithiothreitol (DTT; 1,4-dimercaptobutane-2,3-diol) (0.50 mL) for 15 min, and the waste was discarded. The samples were also alkylated with 0.10 M NH_4HCO_3 (pH 8.1) containing 10 mM EDTA and 0.10 M acrylamide (0.50 mL) for 15 min, and the waste was discarded. The gel pieces were washed with 10% (v/v) AcOH containing 50% (v/v) CH_3OH (0.50 mL) for 40 min with intense mixing, and the waste was discarded. The washing step was repeated four times. The gel pieces were then incubated with 0.10 M NH_4HCO_3 (pH 8.1, 0.50 mL) for 10 min, and the waste was discarded. The gel pieces were shrunk with CH_3CN (0.50 mL) for 10 min, and the waste was discarded. The shrunken gel pieces were completely freeze-dried for 15 min, and the dried gel pieces were swollen in 50 μL of trypsin solution [0.10 M NH_4HCO_3 (pH 8.1) containing 2.5 μg proteomics-grade trypsin] then incubated for 15 min at 37 °C. The incubated gel pieces were washed with 0.10 M NH_4HCO_3 (pH 8.1, 0.25 mL) then incubated again in an identical trypsin solution for 12 h at 37 °C. The tryptic digest was extracted from the gel pieces by ultrasonication in 0.10% (v/v) aqueous TFA containing 50% (v/v) CH_3CN for 15 min and then in 0.10% (v/v) aqueous TFA containing 70% (v/v) CH_3CN for 15 min. The extracted fraction containing ~10 μg ovalbumin digest was freeze-dried and stored at below 0 °C. The tryptic digestion samples were subjected to phosphate-affinity chromatography with a pipette microtip column as described below (Section 2.10).

2.10 *Phosphate-affinity chromatography for phosphopeptides with a pipette microtip column*

A chromatographic microtip column for enrichment of phosphopeptides for subsequent MS analysis was prepared as follows (see Fig. 4A). The end (~25 mm) of a GELoader tip (Eppendorf Japan, Tokyo, Japan) was cut off and then the bottom of the microtip was packed with a custom-made sintered plastics filter (Nichiryo, Koshigaya, Japan). The tip was loaded with Phos-tag Toyopearl beads (~10 μL bed of compressed gel) and then another sintered plastics filter was packed on top of the beads. The resulting microtip column (the Phos-tag tip) was attached to a 20-μL micropipettor (for example, a Gilson P-20; Middleton, WI, USA). The plunger of the micropipettor was gently depressed and released during the analytical processes. First, the storage buffer was discarded. To form the zinc(II)-bound Phos-tag Toyopearl beads, a balancing buffer (20 μL) consisting of 10 mM MES–NaOH (pH 6.0), 0.10 M NaCl, 1.0 mM disodium oxalate, and 10 μM $Zn(OAc)_2$ was aspirated into the pipette microtip column and then dispensed; the waste was discarded. The column was washed with a binding/washing buffer (20 μL) consisting of 10 mM MES–NaOH (pH 6.0), 0.10 M NaCl, 1.0 mM disodium oxalate, and 20% (v/v) CH_3CN. This washing operation was repeated three times. Protein tryptic digests (β-casein or ovalbumin, each 10 μg) were dissolved in the binding/washing buffer (20 μL). The sample solution was aspirated into the column, and the plunger of the micropipettor was then depressed and gently released. This pipetting operation was repeated ten times to ensure maximal binding of the phosphopeptides, and then the waste was discarded. Next, the column was washed with the binding/washing buffer (20 μL) and the waste was discarded. This washing operation was repeated five times. To elute the gel-bound peptides, an elution solution (20 μL) consisting of 1.0% (v/v) aqueous TFA was aspirated, dispensed, released, and finally collected as an elution fraction. This eluting operation was repeated five times, and all the fractions were combined and analyzed by matrix-assisted laser desorption/ionization–time of flight (MALDI-TOF) MS. All the operations involved in the pipette microtip column chromatography were performed at room temperature.

2.11 *MALDI-TOF MS analysis*

The elution fraction containing phosphopeptides, obtained by using a Phos-tag tip as described above, was desalted and concentrated by using a ZipTip-C18 (Millipore). MALDI-TOF mass spectra (positive linear mode) were obtained on a Voyager RP-3 BioSpectrometry Work Station (PerSeptive Biosystems; Framingham, MA, USA) equipped with a nitrogen laser (337 nm, 3-ns pulse). A saturated solution of (2*E*)-2-cyano-3-(4-hydroxyphenyl)acrylic acid (10 mg/mL) in an aqueous solution containing 0.10% (v/v) TFA and 33% (v/v) CH_3CN was used as the matrix. The sample solution (2.0 μL) applied to the MS plate was immediately mixed with the matrix solution (1.0 μL), then dried completely. Each mass spectrum was produced by accumulating data from 128 laser shots. Time-to-mass conversion was performed by means of external calibration using $[M + H]^+$ peaks for two peptides, angiotensin I (*m/z* 1297.51) and adrenocorticotropin hormone/corticotropin-like intermediate lobe peptide 7–38 (*m/z* 3660.19).

2.12 *Open-column phosphate-affinity chromatography for phosphoproteins*

The bottom filter of a polypropylene column (Qiagen) was washed with 20% (v/v) propan-2-ol and the column was then loaded with Phos-tag Toyopearl beads [~2.0 mL of 50% (v/v) suspension] to form a 1.0-mL bed of compressed gel. The gel was washed with 2.0 mL of a binding/washing buffer consisting of 0.10 M Tris–AcOH (pH 7.5) and 0.50 M NaOAc to remove the storage buffer. To form the zinc(II)-bound Phos-tag Toyopearl beads, 2.0 mL of a balancing buffer consisting of 0.10 M Tris–AcOH (pH 7.5), 0.50 M NaOAc, and 10 μM $Zn(OAc)_2$ was placed in the column, and the mixture was incubated for 5 min. The column was then washed with the binding/washing buffer (5.0 mL) to remove excess $Zn(OAc)_2$. The cell lysate from the EGF-stimulated cultured cells (0.50 mg proteins dissolved in 0.25 mL of an RIPA buffer) was diluted with 1.0 mL of the binding/washing buffer. The sample solution (1.25 mL) was applied to the gel and then a flow-through fraction containing unbound proteins was collected. Next, the binding/washing buffer (5.0 mL) was loaded, and a washing fraction was collected. The flow-through and washing fractions were combined and used for subsequent analyses. To elute the gel-bound proteins, an elution buffer (5.0 mL) consisting of 0.10 M Tris–AcOH (pH 7.5), 1.0 M NaCl, and 10 mM NaH_2PO_4–NaOH (pH 7.5) was placed in the column. Each fraction was desalted and condensed by using a Microcon YM-10 centrifugal filter unit and then dissolved in distilled water (0.20 mL). The recovery of proteins in each concentrated fraction was estimated by using a Bio-Rad protein assay kit as described above. All the open-column chromatography operations were performed at room temperature.

2.13 *Two-dimensional gel electrophoresis (2-DE)*

The protein samples were purified by using a ReadyPrep 2-D Clean Up Kit. The purified samples were resolved in the loading buffer supplied with an EzApply 2D Kit and then treated with the alkylation solution with the same kit for 10 min. The sample solution (0.10 mg protein, 20 μL) was applied onto an agarGEL (pH 3–10, 25-mm diameter and 75-mm long), and layered 10 μL of 2.0 M urea. The anode buffer was 10 mM H_3PO_4, and the cathode buffer was 0.20 M NaOH. Isoelectric focusing (IEF) was performed using an Atto AE-6540 at 300 V for 210 min. After electrophoresis, the agarGEL was steeped in 2.5% (w/v) trichloroacetic acid for 3 min twice, next in distilled water for 3 min twice, and then in distilled water for more than 2 h. Finally, the agarGEL was steeped in 0.25 M Tris–HCl (pH 6.8) buffer containing 2.5% (w/v) SDS and 0.10% (w/v) bromophenol blue, then immobilized on the top of the SDS-PAGE gel with the same buffer containing 1.0% (w/v) agarose. SDS-PAGE was performed as described above.

2.14 *Staining of the gel with Pro-Q Diamond, SYPRO Ruby, and Coomassie Brilliant Blue*

The SDS-PAGE gel was fixed in an aqueous solution containing 50% (v/v) MeOH and 10% (v/v) AcOH for 30 min twice. To stain the phosphoproteins with Pro-Q Diamond, the fixed gels were washed three times in distilled water for 30 min each time, then incubated with Pro-Q Diamond phosphoprotein gel stain for 3 h, and finally washed in 50 mM NaOAc–AcOH (pH 4.0) buffer containing 20% (v/v) CH_3CN for 3 h. For SYPRO Ruby staining, fixed gels or Pro-Q Diamond-stained gels were incubated with SYPRO Ruby protein gel stain for 2 h then washed in 10% (v/v) MeOH and 7.0% (v/v) AcOH for 2 h. For Coomassie Brilliant Blue (CBB) staining, fixed gels or SYPRO Ruby-stained gels were incubated with a CBB solution [0.10% (w/v) CBB, 10% (v/v) AcOH, and 40% (v/v) MeOH] for 1 h, and then washed in an aqueous solution containing 10% (v/v) MeOH and 7.0% (v/v) AcOH until the background was clear. Fluorescence images of the gel were acquired on an FLA 5000 laser scanner (Fujifilm). An LAS 3000 image analyzer was used for the observation of CBB-stained gel.

3. Results and Discussion

3.1 *Affinity spin column chromatography for phosphopeptides*

To determine the optimal binding/washing conditions for the separation of peptides in phosphate-affinity chromatography using Phos-tag Toyopearl, we adopted a spin column method (50-μL gel scale) using twelve commercially available neutral, acidic, and basic peptides, some of which were phosphorylated; these are listed in Table 1.

Table 2 shows the effects of the pH of the binding/washing buffer and of additional salts in the buffer on separation of phosphopeptides and nonphosphopeptides. A mixture containing 5.0 nmol each of [Ala-97]-RII 81–99 (No. 4), RII phosphopeptide (No. 5), PKC substrate peptide (No. 7), phosphorylated PKC substrate (No. 8), and 4A/4B peptide (No. 9) was applied. After spin column chromatography, the relative recovery of sample peptides in each fraction was determined by reversed-phase HPLC analysis. Some binding/washing buffer conditions were examined. We selected buffer pH values of 7.4 (0.10 M Tris–AcOH) and 6.0 (10 mM

Table 1. Amino acid sequences of the phosphopeptides and nonphosphopeptides. Phosphorylated amino acids are underlined.

	No.	Peptide Name	Amino Acid Sequence
Neutral Peptides	1	Insulin receptor 1142–1153	TRDIYETDYYRK
	2	Kinase domain of insulin receptor-2	TRDIpYETDYYRK
	3	Kinase domain of insulin receptor-5	TRDIpYETDpYpYRK
	4	[Ala-97] RII 81–99	DLDVPIPGRFDRRVSVAAE
	5	RII phosphopeptide	DLDVPIPGRFDRRVpSVAAE
	6	MAP 177–189 pT/pY peptide	DHTGFLpTEpYVATR
Basic Peptides	7	PKC substrate peptide	QKRPSQRSKYL
	8	Phosphorylated PKC substrate	KRPpSQRHGSKY-NH_2
Acidic Peptides	9	4A/4B peptide	DEMEECSQHLPNI
	10	Src peptide 521–533	TSTEPQYQPGENL
	11	Phosphorylated Src peptide 521–533	TSTEPQpYQPGENL
	12	EGFR 988–998	DADEpYQPGENL

Table 2. Effects of pH values and additional salts in the binding/washing processes on the separation of phosphopeptides and nonphosphopeptides by affinity spin column chromatography using Phos-tag Toyopearl.

(A)		Peptides[#]				
	Washing buffer (pH, additional salts)	Neutral No. 4	No. 5	Basic No. 7	No. 8	Acidic No. 9
(a)	pH 7.4 + 0.10 M NaCl	1%	82%	0%	52%	46%
(b)	pH 7.4 + 0.10 M NaOAc	1%	74%	0%	52%	21%
(c)	pH 6.0 + 0.10 M NaCl	1%	91%	1%	88%	34%
(d)	pH 6.0 + 0.10 M NaOAc	2%	85%	1%	74%	3%

(B)		Peptides[#]				
	Washing buffer	Neutral No. 4	No. 5	Basic No. 7	No. 8	Acidic No. 9
(a)	+ 0.10 M F_3CCO_2Na	3%	91%	11%	87%	57%
(b)	+ 0.10 M NaOAc	2%	92%	11%	88%	67%
(c)	+ 5.0 mM $(COONa)_2$	1%	73%	99%	100%	1%
(d)	+ 0.10 M NaCl and 1.0 mM $(COONa)_2$	1%	88%	1%	88%	4%

MES–NaOH). To prevent nonspecific binding to the phosphate-affinity beads, 0.10 M NaCl or 0.10 M NaOAc was added to the buffer solutions. The separation results in terms of the percentage recovery of each peptide in the elution fractions are shown in Table 2A. In buffer systems with a pH of 6.0 (entries c and d), the elution fraction contained a greater proportion of phosphopeptides (Nos. 5 and 8) and a smaller proportion of nonphosphopeptides (Nos. 4, 7, and 9) than did the buffer systems with a pH of 7.4 (entries a and b). In particular, the nonspecific binding of acidic peptide (No. 9) decreased to only 3% in a buffer system of pH 6.0 containing NaOAc (entry d), and the separation was markedly improved. The acidic peptides are in many cases the most critical factor in the enrichment of phosphopeptides. In this case, the presence of an excess of acetate anion and lower pH should competitively inhibit nonspecific binding between the carboxylate groups of the peptides and Phos-tag. To observe the effects of other additional salts, F_3CCO_2Na, NaOAc, and disodium oxalate were added to the 10 mM MES–NaOH buffer (pH 6.0). The separation results in terms of the percentage recovery of each peptide in the elution fractions are shown in Table 2B. In buffer systems containing 0.10 M F_3CCO_2Na (entry a) or 0.10 M NaOAc (entry b), a considerable degree of nonspecific binding of peptide No. 9 was observed in the elution fractions (57% and 67%, respectively). When the concentration of F_3CCO_2Na was increased to 0.25 M, specific binding of phosphopeptides was decreased (data not shown). In the buffer system containing disodium oxalate (entry c), an oxalate concentration of the order of only mM (5.0 mM of disodium oxalate) had a marked effect in preventing nonspecific binding of peptide No. 9, but had no effect on peptide No. 7. In the buffer system that contained both 1.0 mM disodium oxalate and 0.10 M NaCl (entry d), the separation was improved compared with that in the 10 mM MES–NaOH buffer (pH 6.0) containing 0.10 M NaOAc (see entry d in Table 2A). These findings showed that the most effective binding/washing buffer system for the separation of phosphopeptides by phosphate-affinity chromatography on Phos-tag Toyopearl in this model study consisted of 10 mM MES–NaOH (pH 6.0) containing 0.10 M NaCl and 1.0 mM disodium oxalate.

Next, we examined the optimal elution conditions for phosphate-affinity chromatography using the same spin column method and a triphosphorylated peptide, kinase domain of insulin receptor-5 (No. 3, Table 1). A higher concentration of phosphate in the elution buffer was required to elute the triphosphorylated peptide. The relative recovery of peptide No. 3 in each fraction from spin column chromatography was determined by reversed-phase HPLC analysis. Some eluent systems were found to be suitable for separation of phosphopeptides on Phos-tag Toyopearl. One was 0.10 M to 0.20 M NaH_2PO_4–NaOH (pH 7.4) as a neutral pH eluent. Others were volatile acidic solutions, such as aqueous 0.50% (v/v) TFA, or basic solutions, such as aqueous 2.0 M NH_3. Furthermore, we succeeded in separating a triphosphopeptide from a monophosphopeptide (kinase domain of insulin receptor-2; peptide No. 2, Table 1) present in 5.0 nmol amounts by stepwise elution with two kinds of phosphate buffer (Fig. 2). Elution was conducted with 1.0 mM NaH_2PO_4–NaOH (pH 7.4) (repeated six times, fractions 1–6) and with 0.10 M NaH_2PO_4–NaOH (pH 7.4) (repeated four times, fractions 7–10). Neither the flow-through fraction nor the washing fraction contained both phosphopeptides. Fractions 1–6 contained almost all the monophosphopeptide (recovery 99%, purity 96%) and fractions 7–10 contained almost the triphosphopeptide (recovery 96%, purity 99%). The total time required to complete the procedure shown in Fig. 2 was less than 30 min.

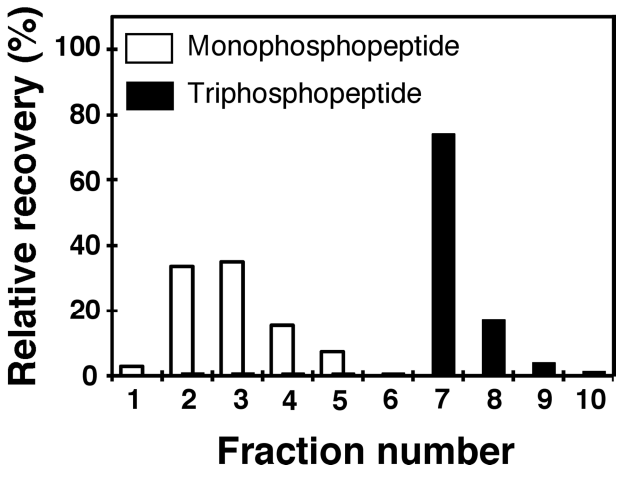


Figure 2. Relative recoveries of a monophosphopeptide (kinase domain of insulin receptor-2, see entry 2 in Table 1) and of a triphosphopeptide (kinase domain of insulin receptor-5, see entry 3 in Table 1) by affinity spin column chromatography using Phos-tag Toyopearl. Stepwise elution with NaH_2PO_4–NaOH (pH 7.4) from 1.0 mM (fractions Nos. 1–6) to 0.10 M (Nos. 7–10) was performed. Open and closed bars represent the recovery values of mono- and triphosphorylated peptide, respectively.

3.2 *Comparison of novel Zn(II)-IMAC using Phos-tag Toyopearl with other methods*

To evaluate the performance of the novel Zn(II)-IMAC using Phos-tag Toyopearl, we compared Phos-tag-based phosphate-affinity chromatography with other existing methods in the analysis of a sample mixture of twelve peptides (5.0 nmol of each peptide), as described above (Table 1). Phos-tag Agarose (Wako Pure Chemical), PHOS-select Iron Affinity Gel (Sigma-Aldrich), PhosphoProtein Purification Kit (Qiagen), and Titansphere TiO (GL Sciences) were used as objects for comparison, and we adopted the spin column procedure (50-µL gel scale). Figure 3 shows HPLC chromatograms for the mixture of peptides before separation (upper), for the flow-through/washing fractions (center), and for the eluted fractions (lower). The underlined numbers correspond to phosphopeptides. The results for the separation of phosphopeptides by the novel Zn(II)-IMAC using Phos-tag Toyopearl are shown in Fig. 3A. In the chromatographic separation, a binding/washing buffer consisting of 10 mM MES–NaOH (pH 6.0) containing 0.10 M NaH_2PO_4–NaOH (pH 7.4). In the flow- through/washing fraction, phosphopeptides No. 5 and No. 8 showed slight leakages, but all the nonphosphopeptides were washed out from the column, and only phosphopeptides were observed in the eluted fraction. The recovery of phosphopeptides was nearly 100% except in the cases of No. 5 and No. 8.

The results obtained by using Phos-tag Agarose areshown in Fig. 3B. A binding/washing buffer of 10 mM MES–NaOH (pH 6.0) containing 0.10 M NaCl and 5.0 mM disodium oxalate and an elution buffer consisting of 0.10 M NaH_2PO_4–NaOH (pH 7.4) were used. A higher concentration of oxalate in the binding/washing buffer was required to perform quality chromatography. The chromatograms are almost the same as those shown in Fig. 3A. Therefore, in the separation of phosphopeptides, Phos-tag Agarose provided a degree of selection similar to that of Phos-tag Toyopearl.

The results obtained by using the Sigma-Aldrich beads are shown in Fig. 3C. These beads consist of a matrix of agarose on which the active metal center is mono-Fe(III). A binding/washing buffer of 0.25 M AcOH (pH 2.7) containing 30% (v/v) CH_3CN, and an elution buffer of 0.20 M NaH_2PO_4–NaOH (pH 7.4) were used according to the standard protocol suggested by Sigma-Aldrich. The elution fraction contained not only phosphopeptides but also a small amount of the acidic peptide No. 9 (7%). The relative recovery of phosphopeptides (49% of No. 11, 88% of No. 2, and 70% of No. 12) was inferior to that achieved with Zn(II)-IMAC using Phos-tag Toyopearl. In addition, this method requires acidic pH conditions in the binding/washing process, and is therefore not suitable for the separation of native phosphoproteins.

The results achieved with the Qiagen beads are shown in Fig. 3D. The binding/washing buffer (pH 6.0) and the elution buffer (pH 7.4) supplied with the kit were used. In the elution fraction, phosphopeptides were observed exclusively. However, the relative recovery of phosphopeptides was inferior to that achieved with the Phos-tag Toyopearl beads.

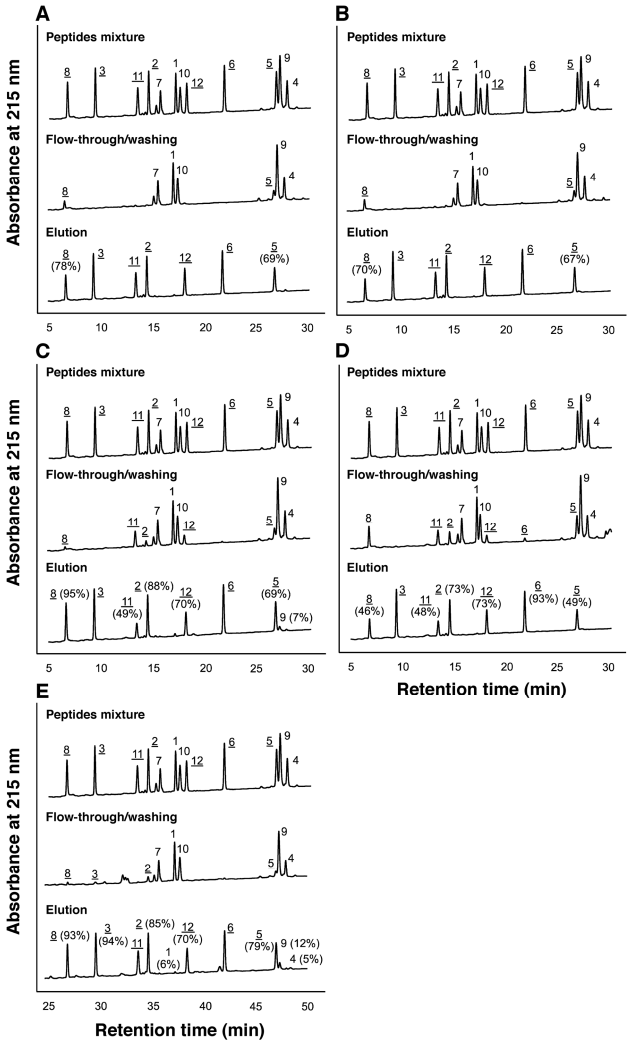


Figure 3. Comparison with other existing techniques. Phosphopeptides were separated from a mixture of twelve peptides (see Table 1) by affinity spin column chromatography using Phos-tag Toyopearl (A), Phos-tag Agarose (B), PHOS-select Iron Affinity Gel (C), PhosphoProtein Purification Kit (D), and Titansphere TiO (E). The peptides mixture prior to loading on each resin (upper chromatograms), in the flow-through/washing fraction (middle chromatograms), and in the elution fractions (lower chromatograms) were analyzed by reversed-phase HPLC. The underlined numbers correspond to phosphopeptides.

The results obtained with GL Sciences beads are shown in Fig. 3E. The beads consist of spherical particles of titanium dioxide (TiO_2). A binding/washing solution of aqueous 0.10% (v/v) TFA containing 0.30 g/mL lactic acid and 80% (v/v) CH_3CN, and both solutions of 0.20 M NaH_2PO_4–NaOH (pH 7.4) and aqueous 2.0 M NH_3 were used for elution in accord with a previous report [11]. Phosphopeptides were found in the eluent, together with small amounts of nonphosphopeptides (6% of No. 1, 12% of No. 9, and 5% of No. 4). The relative recovery of phosphopeptides was excellent. This method requires an acidic pH during the binding/washing process (pH 1.3) and, like the method using the PHOS-select Iron Affinity Gel, cannot therefore be used to separate native phosphoproteins. Furthermore, this method suffers from a

problem in the HPLC analysis: Because the flow-through/washing fraction (the high viscosity of the mixture) contained large amounts of lactic acid, the HPLC elution process of the acid with a eluent consist of aqueous 0.10% (v/v) TFA and 1.0% (v/v) CH_3CN for 20 min was necessary before the peptide analysis as described in the Methods section.

We have therefore demonstrated that our new method for analyzing phosphopeptides by using Phos-tag Toyopearl has advantages over other existing methods.

3.3 *Affinity pipette microtip column chromatography for phosphopeptides*

We developed a pipette microtip column filled with the Phos-tag Toyopearl beads (Phos-tag tip; 10-μL gel scale, see Fig. 4A), and we demonstrated its use in the separation of small amounts of phosphopeptides from a protein tryptic digest for subsequent MALDI-TOF/MS analysis. The tryptic digest of β-casein and in-gel tryptic digest of ovalbumin were used as the first model samples (each 10 μg). A binding/washing buffer of 10 mM MES–NaOH (pH 6.0) containing 0.10 M NaCl, 1.0 mM disodium oxalate, and 20% (v/v) CH_3CN, and an elution solution of 1.0% (v/v) aqueous TFA were used in the chromatographic process. The total time required to complete the microchip column chromatography and desalting procedures was less than 30 min.

Two phosphopeptides were obtained from the tryptic digest of β-casein. One was a monophosphopeptide consisting of 16 residues (Phe_{33}-Gln-*pSer*-Glu-Glu-Gln-Gln-Gln-Thr-Glu-Asp-Glu-Leu-Gln-Asp-Lys_{48}, P1), and the other was a tetraphosphopeptide consisting of 25 residues (Arg_{16}-Glu-Leu-Glu-Glu-Leu-Asn-Val-Pro-Gly-Glu-Ile-Val-Glu-*pSer*-Leu-*pSer*-*pSer*-*pSer*-Glu-Glu-Ser-Ile-Thr-Arg_{40}, P2). In the mass spectrum of the tryptic digest of β-casein before the separation, no peaks for the phosphopeptides were detected (data not shown). Figure 4B shows a typical mass spectrum of the elution fraction (β-casein). We detected two peaks and we were able to identify the P1 monophosphopeptide (33–48; m/z = 2062.9) and P2 tetraphosphopeptide (16–40; m/z = 3124.8).

Similarly, we obtained two monophosphopeptides from the in-gel tryptic digest of ovalbumin separated by SDS-PAGE. One consisted of 23 residues (Leu_{62}-Pro-Gly-Phe-Gly-Asp-*pSer*-Ile-Glu-Ala-Gln-Cys-Gly-Thr-Ser-Val-Asn-Val-His-Ser-Ser-Leu-Arg_{84}) and the other consisted of 20 residues (Glu_{340}-Val-Val-Gly-*pSer*-Ala-Glu-Ala-Gly-Val-Asp-Ala-Ala-Ser-Val-Ser-Glu-Glu-Phe-Arg_{359}). The Cys_{73} residue was alkylated with acrylamide during the process of in-gel digestion. In the mass spectrum of the tryptic digest of ovalbumin, no peaks corresponding to phosphopeptides were detected (data not shown). Figure 4C shows a typical mass spectrum of the elution fraction (ovalbumin). We detected three distinct peaks, and we were able to identify a monophosphopeptide (P3, 340–359, m/z = 2089.1), a second monophosphopeptide (P4, 62–84, m/z = 2526.1), and a nondigested monophosphopeptide (P5, 59–84, m/z = 2916.9).

We succeeded in separating and identifying phosphopeptides from the peptides mixture by using the Phos-tag tip without detecting any nonphosphopeptides. The advantages of Phos-tag tip were that it requires only simple operation of pipetting without any special devises, and that it has a short operational time. In addition, the novel Phos-tag Toyopearl beads are much more stable during long-term storage than are Phos-tag Agarose biopolymer beads, and they can be reused many times for separation of low-molecular weight phosphorylated compounds containing phosphopeptides. We confirmed that previously used Phos-tag Toyopearl beads that had been stored in the binding/washing buffer for 6 months at room temperature had the same potency in separation as beads prepared shortly before use (data not shown).

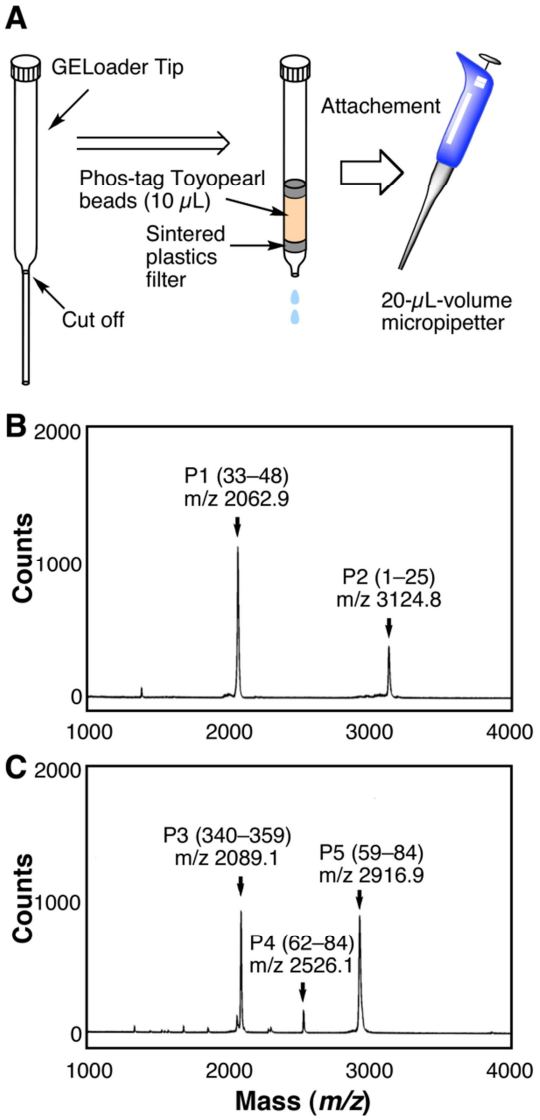


Figure 4. Affinity pipette microtip column chromatography for the separation of phosphopeptides. Scheme for the preparation of the pipette microtip column packed with Phos-tag Toyopearl (Phos-tag tip) (A). MALDI-TOF/MS analysis of each elution fraction after the affinity pipette microtip column chromatographic separation of the tryptic digest of β-casein (B) and in-gel tryptic digest of ovalbumin (C) as samples.

3.4 *Optimization of buffers for the separation of phosphoproteins*

We extended the application of the Phos-tag Toyopearl beads to the separation of phosphoproteins. In our previous study on the separation of native phosphoproteins from cell lysates by using Phos-tag Agarose, we conducted the binding/washing process with 0.10 M Tris–AcOH (pH 7.5) containing 1.0 M NaOAc, and the elution process with 0.10 M Tris–AcOH (pH 7.5) containing 10 mM NaH_2PO_4–NaOH (pH 7.5) and 1.0 M NaCl [16]. During the Zn(II)-IMAC process, no detergent or reducing agent was used. In addition, we have proposed that appropriate concentration of NaOAc (1.0 M) should be added to the binding/washing buffer to avoid nonspecific binding of proteins to the Phos-tag Agarose biopolymer matrix. We applied a similar buffer system with the Phos-tag Toyopearl beads and, in particular, we performed a detailed optimization of the concentration of NaOAc in the binding/washing buffer. We used an EGF-stimulated A431 cell lysate as a real biological sample, and we adopted the spin column method (50-μL gel scale). After elution, the residual proteins remaining in the column were analyzed by washing the column with a buffer containing 0.10 M Tris–AcOH, 1.0% (w/v) SDS, and 10 mM EDTA. Figure 5 shows typical results for SDS-PAGE with SYPRO Ruby gel staining and with Western blotting for the flow-through/washing fraction (lane 1), elution fraction (lane 2), and column-washing fraction (lane 3) for a binding/washing buffer containing 0, 0.25, 0.50, or 1.0 M NaOAc. The SYPRO Ruby images (Fig. 5A) show the distribution of proteins in each fraction. In the absence of the NaOAc, there were scarcely any proteins in the flow-through/washing fraction; in other words, almost all the proteins, including nonphosphorylated ones, were bound to the column. In addition, many proteins remained on the column after elution. As the concentration of NaOAc was increased, fewer proteins were left in the column after elution, and many proteins were distributed in the flow-through/washing and elution fractions.

The efficiency of the separation of phosphoproteins was determined by means of Western blotting using the complex of Phos-tag Biotin with HRP-conjugated streptavidin (Fig. 5B). Probing with Phos-tag Biotin permitted the comprehensive detections of phosphoproteins as ECL signals (the single positive control band corresponding to the standard phosphoprotein, ovalbumin, can be seen in the molecular-weight markers lane M in Fig. 5B). In the presence of less than 0.25 M NaOAc, ECL signals were detected from both the elution and column-washing fractions. On the other hand, in the presence of 0.50 or 1.0 M NaOAc, the signals were detected from the elution fraction only. We therefore decided to use a 0.50 M concentration of NaOAc in the binding/washing buffer for this model study. Under the optimized conditions, we also examined the separation of the phosphorylated and nonphosphorylated MAPK1/2 proteins (Figs. 5C and 5D). It is known that MAPK1/2 (Erk1/2) is phosphorylated from EGF signaling. The strong ECL signal obtained by probing with an anti-pMAPK1/2 antibody was detected in the elution fraction

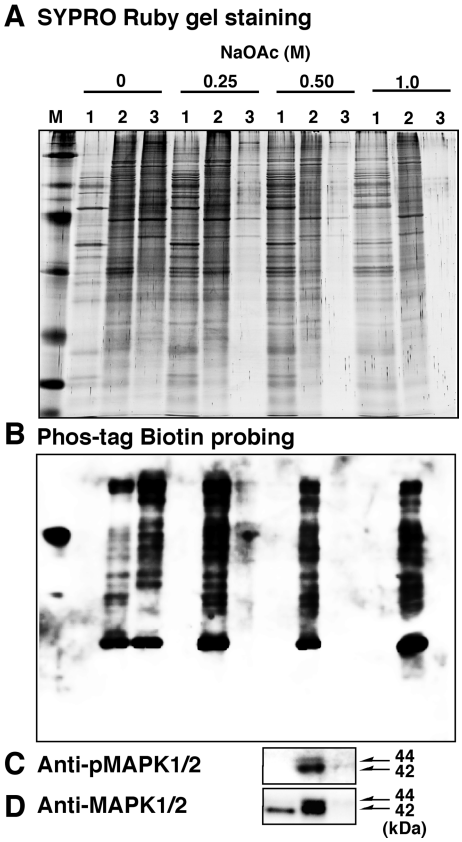


Figure 5. Optimization of the concentration of NaOAc in binding/washing buffer on the separation of phosphoproteins in an EGF-stimulated A431 cell lysate. Concentrations of 0, 0.25, 0.50, and 1.0 M NaOAc were used in the binding/washing processes. The flow-through/washing, elution, and column-washing fractions were applied in lanes 1, 2, and 3, respectively. The SDS-PAGE gels were analyzed by SYPRO Ruby gel staining (A) or by Western blotting with Phos-tag Biotin (B). Molecular-weight standards for 97, 66, 45 (standard phosphoprotein, ovalbumin), 29, 20, and 14 kDa are shown, in order from the top, in lane M. The fractions collected when 0.50 M NaOAc was used were analyzed by Western blotting with anti-pMAPK1/2 (C) and anti-MAPK1/2 (D) antibodies.

exclusively (C). In contrast, the signal obtained by probing with an anti-MAPK1/2 antibody was detected in both the flow-through/washing and elution fractions (D). These results of Western blotting indicated that phosphorylated proteins were separated from their corresponding nonphosphorylated counterparts by the Phos-tag Toyopearl beads.

3.5 *Enrichment of phosphoproteins using an open-column method*

To evaluate the enrichment of phosphoproteins in the elution fraction, we initially performed phosphate-affinity chromatography using a comparative large amount of the cell lysate and an open column with the Phos-tag Toyopearl beads (1-mL gel scale). An EGF-stimulated A431 cell lysate (0.50 mg solubilized cellular proteins in 0.25 mL of an RIPA buffer) was prepared from the cultured cells and then diluted with 1.0 mL of the binding/washing buffer. The resulting solution (1.25 mL) was loaded onto the open column. Details

of the procedures adopted for affinity column chromatography are described in the Methods section. The total time required for the phosphate-affinity chromatography procedure was less than 40 min. For SDS-PAGE followed by Western blotting (Fig. 6), fractions of the cell lysate before chromatography (lane 1) and after elution (lane 2), and the flow-through/washings (lane 3) (6.0 μg of proteins per lane) were sequentially applied. The CBB-staining image of the blotted PVDF membrane (A) showed that the amounts of proteins in each lane were almost equal. Next, we determined the distribution of the various phosphoproteins by Western blotting using Phos-tag Biotin (B), anti-pSer antibody (C), anti-pMAPK substrates (PXTP) antibody (D), anti-pMAPK1/2 antibody (E), anti-pShc antibody (F), and anti-pErbB-2/HER-2 antibody (G). Shc and ErbB-2/HER-2, as well as MAPK1/2, are candidates for *in vivo* phosphorylation in EGF signaling. In all blottings, the ECL signals corresponding to phosphoproteins were stronger in the elution fraction (lane 2) than in the lysate (lane 1). There were very small amounts of phosphoprotein in the flow-through/washing fractions (lane 3 of B and C). These results confirmed that the cellular phosphoproteins are strongly enriched in the elution fraction.

We quantified the proteins in each fraction by using 0.25 mL of an RIPA buffer containing 0.50 mg protein of the EGF-stimulated A431 cell lysate. The amounts of recovered protein in the flow-through/washing and elution fractions were 50% and 22%, respectively. The yield of the elution fraction is consistent with the general content of phosphoproteins in mammalian cells (10–30%) [27]. When larger amounts of the lysate proteins (e.g., 0.60 mg in 0.25 mL of an RIPA buffer) were loaded into the column, some of the phosphoproteins were eluted in the flow-through/washing fraction (data not shown). A similar leakage into the flow-through/washing fraction resulted from the use of twice the volume of an RIPA buffer (0.50 mL) containing the solubilized proteins (0.50 mg), possibly as a result of competitive binding of $HOVO_3^{2-}$ or of elimination of zinc(II) by EDTA. Thus, the appropriate capacity of the 1 mL-compressed Phos-tag Toyopearl column is ~0.50 mg of lysed proteins in 0.25 mL of RIPA buffer. These results that we obtained by using the Phos-tag Toyopearl beads were almost identical to those that we obtained by using Phos-tag Agarose, as reported previously [16].

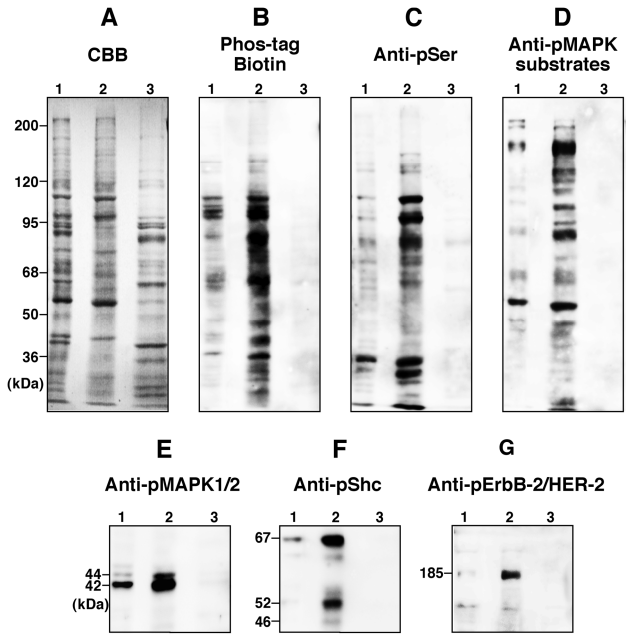


Figure 6. Comparison of the relative amounts of phosphoproteins from an EGF-stimulated A431 cell lysate before and after affinity open column chromatography. The cell lysate before loading on the column (lane 1), the elution fraction (lane 2), and flow-through/washing fraction (lane 3) were subjected to SDS-PAGE. Each lane contained 6.0 μg proteins. The proteins on the SDS-PAGE gels were electrotransferred to PVDF membranes and visualized by CBB gel staining (A) or by Western blotting using Phos-tag Biotin (B), anti-pSer antibody (C), anti-pMAPK substrates (PTXT) antibody (D), anti-pMAPK1/2 antibody (E), anti-pShc antibody (F), or anti-pErbB-2/HER-2 antibody (G). The molecular weights are shown on the left-hand side.

Finally, we demonstrated the value of phosphate-affinity chromatography with Phos-tag Toyopearl as an efficient procedure for the separation and enrichment of phosphoproteins by examining the 2-DE (IEF and SDS-PAGE) of each fraction obtained from the open column with 1 mL of gel, followed by total protein gel staining with SYPRO Ruby, phosphoprotein gel staining with Pro-Q Diamond, Western blotting, and MS analysis (Fig. 7). The cell lysate before the chromatography (left panels), after elution (center panels), and after flow-through/washing (right panels) (100 μg of proteins each) were separated on a 2-DE gel that was then stained with Pro-Q Diamond (B). In subsequent SYPRO Ruby gel staining (A), many spots that were not detected by Pro-Q Diamond staining were observed in the cell lysate and in both the elution and flow-through/washing fractions. That is partially explained by the fact that the cell lysate and flow-through/washing fraction contain many nonphosphorylated proteins. However, SYPRO Ruby staining of the elution fraction permitted the detection of many protein spots, especially in alkaline isoelectric point (p*I*) region, that were not detected by Pro-Q Diamond staining. We next performed Western blotting analysis by treatment with the anti-pSer antibody (C). Some protein spots in the alkaline p*I* region that showed no fluorescent signal with Pro-Q Diamond were detected by probing with the anti-pSer antibody. We noticed that the Pro-Q Diamond stain tended to detect phosphoproteins in region of acidic p*I* much more strongly than it did in region of alkaline p*I*. Furthermore, we identified 10 proteins in the SYPRO Ruby-stained gel (Nos. 1–10, see center panel of A) by LC-MS/MS analysis performed by a commercial contract research company, Applied Cell Biotechnologies (Yokohama, Japan). The results were examined by using the Mascot search engine (Matrix Science, London, UK). All the proteins were well-known phosphoproteins, as summarized in Table 3. In addition, all the molecular weights and p*I* values were consistent with the location of the corresponding protein spot on the 2-DE gel. These results confirm that phosphoprote-

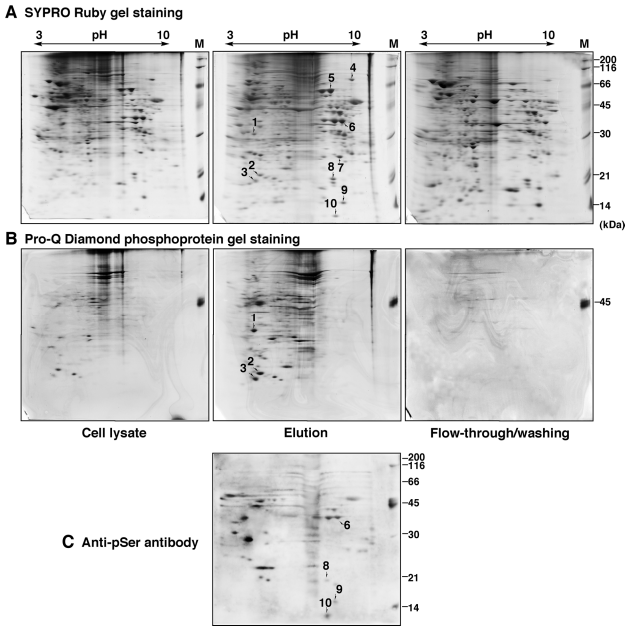


Figure 7. 2-DE (IEF and SDS-PAGE) analysis of an EGF-stimulated A431 cell lysate before and after affinity open column chromatography. The cell lysate before loading on the column (left), the elution fraction (center), and the flow-through/washing fraction (right) were subjected to 2-DE. Each 2-D gel contains 100 μg proteins. The gels were stained with SYPRO Ruby gel stain (A) and Pro-Q Diamond phosphoprotein gel stain (B). The elution fraction was probed by Western blotting analysis using anti-pSer antibody (C). The molecular-weight standards for 200, 116, 66, 45 (standard phosphoprotein, ovalbumin), 30, 21, and 14 kDa are shown, in order from the top, in lane M.

omics could progress markedly if our phosphate-affinity chromatography were combined with existing methods involving 2-DE and recent advances in MS analysis.

4. Concluding remarks

We have demonstrated a simple and efficient method for the separation and enrichment of phosphopeptides and phosphoproteins by means of Zn(II)–IMAC with a novel phosphate capture bead, Phos-tag Toyopearl. Phos-tag Toyopearl (15 μmol/mL-gel of Phos-tag) was synthesized by coupling a primary amine derivative of Phos-tag with NHS-activated Toyopearl AF-Carboxy-650M gel. Although there are other well-known types of IMAC that use other metal ions such as Fe(III) or Ga(III) and types of MOAC that use metal oxides such as titanium dioxide (TiO_2) or zirconium dioxide (ZrO_2), neither group of techniques functions quite satisfactorily in phosphoproteomics in terms of their selectivity and versatility. The method based on the metal oxide TiO_2 has been shown to be the best among these techniques [11,13]; however, to perform high-quality chromatography, it is necessary to use an acidic solution (pH <2) containing a high concentration of lactic acid in the binding/washing process. Our Zn(II)-IMAC provides a simple, rapid, and specific procedure for separating phosphopeptides and phosphoproteins in an aqueous solution, and physiological pH values are maintained throughout the analysis. In addition, column of Phos-tag Toyopearl are stable to long-term storage and can be reused many times. Because the procedure is nondenaturing and maintains the conformation and activity of the proteins, the separated phosphoproteins are ideal for use in many downstream applications and in top-down proteomics strategies. Quantitative phosphoproteomics could progress greatly if our new method were combined with the iTRAQ technology [28] or with Phos-tag SDS-PAGE methodology [29–37], as reported previously. The new separation and enrichment procedure is bound to lead to an increase in the sensitivity of evaluation of protein phosphorylation involving the characterization of multiple alternative forms of heterogeneous phosphoprotein, detection of activities of kinases and phosphatases, and elucidation of interactions among proteins, and it should provide an opportunity for a variety of types of experiment.

Acknowledgements

We wish to thank the Research Center for Molecular Medicine and the Analysis Center of Life Science, Hiroshima University, Japan, for the use of their facilities. This work was supported in part by Grants-in-Aid for Scientific Research (B, 22390006; C, 22590037) from the Japan Society for the Promotion of Science (JSPS), by a Grant-in Aid for Young Scientists (B, 22790037) from the Ministry of Education Culture, Sports, Science, and Technology (MEXT), by research grants for Feasibility Study from the Japan Science and Technology Agency (JST), and by a research grant from Chugoku Industrial Innovation Center. MANAC Incorporated, Fukuyama, Japan, also provided some financial support.

References

1. T. Hunter, Curr. Opin. Cell Biol. 21 (2009) 140–146.
2. A. Pandey, A.V. Podtelejnikov, B. Blagoev, X.R. Bustelo, M. Mann, H.F. Lodish, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 179–184.
3. M. Grønborg, T.Z. Kristiansen, A. Stensballe, J.S. Andersen, O. Ohara, M. Mann, O.N. Jensen, A. Pandey, Mol. Cell. Proteomics 1 (2002) 517–527.
4. M. Matsumoto, K. Oyamada, H. Takahashi, T. Sato, S. Hatakeyama, K.I. Nakayama, Proteomics 9 (2009) 3549–3563.
5. L. Andersson, J. Porath, Anal. Biochem. 154 (1986) 250–254.
6. A. Sano, H. Nakamura, Anal. Sci. 20 (2004) 861–864.
7. S.B. Ficarro, M.L. McCleland, P.T. Stukenberg, D.J. Burke, M.M. Ross, J. Shabanowitz, D.F. Hunt, F.M. White, Nat. Biotechnol. 20 (2002) 301–305.
8. M.R. Larsen, T.E. Thingholm, O.N. Jensen, P. Roepstorff, T.J.D. Jørgensen, Mol. Cell. Proteomics 4 (2005) 873–886.
9. M. Kokubu, Y. Ishihama, T. Sato, T. Nagasu, Y. Oda, Anal. Chem. 77 (2005) 5144–5154.
10. H.K. Kweon, K. Håkansson, Anal. Chem. 78 (2006) 1743–1749.
11. N. Sugiyama, T. Masuda, K. Shinoda, A. Nakamura, M. Tomita, Y. Ishihama, Mol. Cell. Proteomics 6 (2007) 1103–1109.
12. X. Zhang, J. Ye, O.N. Jensen, P. Roepstorff, Mol. Cell. Prote-

omics 6 (2007) 2032–2042.
13. Y. Kyono, N. Sugiyama, K. Imami, M. Tomita, Y. Ishihama, J. Proteome Res. 7 (2008) 4585–4593.
14. E. Kinoshita, A. Yamada, H, Takeda, E. Kinoshita-Kikuta, T, Koike, J. Sep. Sci. 28 (2005) 155–162.
15. E. Kinoshita, M. Takahashi, H. Takeda, M. Shiro, T. Koike, Dalton Trans. (2004) 1189–1193.
16. E. Kinoshita-Kikuta, E. Kinoshita, A. Yamada, M. Endo, T. Koike, Proteomics 6 (2006) 5088–5095.
17. E. Kinoshita-Kikuta, E. Kinoshita, T. Koike, Anal. Biochem. 389 (2009) 83–85.
18. T. Takahashi, T. Shiyama, T. Mori, K. Hosoya, A. Tanaka, Anal. Bioanal. Chem. 385 (2006) 122–127.
19. H. Daxecker, M. Raab, E. Bernard, M. Devocelle, A. Treumann, N. Moran, Anal. Biochem. 374 (2008) 203–212.
20. Y. Kato, K. Nakamura, T. Hashimoto, J. Chromatogr. 253 (1982) 219–225.
21. Y. Kato, T. Kitamura, T. Hashimoto, J. Chromatogr. 268 (1983) 425–436.
22. K. Shimura, K. Kasai, S. Ishii, J. Chromatogr. 350 (1985) 265–272.
23. K. Inamori, M. Kyo, Y. Nishiya, Y. Inoue, T. Sonoda, E. Kinoshita, T. Koike, Y. Katayama, Anal. Chem. 77 (2005) 3979–3985.
24. E. Kinoshita, E. Kinoshita-Kikuta, K. Takiyama, T. Koike, Mol. Cell. Proteomics 5 (2006) 749–757.
25. T. Nakanishi, E. Ando, M. Furuta, E. Kinoshita, E. Kinoshita-Kikuta, T. Koike, S. Tsunasawa, O. Nishimura, J. Biomol. Tech. 18 (2007) 278–286.
26. U.K. Laemmli, Nature (London, U. K.) 227 (1970) 680–685.
27. M. Mann, S.E. Ong, M. Grønborg, H. Steen, O.N. Jensen, A. Pande, Trends Biotechnol. 20 (2002) 261–268.
28. T. Nabetani, Y.J. Kim, M. Watanabe, Y. Ohashi, H. Kamiguchi, Y. Hirabayashi, Proteomics 9 (2009) 5525–5533.
29. E. Kinoshita-Kikuta, Y. Aoki, E. Kinoshita, T. Koike, Mol. Cell. Proteomics 6 (2007) 356–366.
30. K. Takeya, K. Loutzenhiser, M. Shiraishi, R. Loutzenhiser, M.P. Walsh, Am. J. Physiol. Renal Physiol. 294 (2008) F1487–F1492.
31. A.E. Messer, C.E. Gallon, W.J. McKenna, C.G. Dos Remedios, S.B. Marston, Proteomics Clin. Appl. 3 (2009) 1371–1382.
32. E. Kinoshita, E. Kinoshita-Kikuta, H. Ujihara, T. Koike, Proteomics 9 (2009) 4098–4101.
33. E. Kinoshita, E. Kinoshita-Kikuta, T. Koike, Nat. Protoc. 4 (2009) 1513–1521.
34. S. Deswal, K. Beck-García, B. Blumenthal, E.P. Dopfer, W.W. Schamel, Immunol. Lett. 130 (2010) 51–56.
35. T. Hosokawa, T. Saito, A. Asada, K. Fukunaga, S. Hisanaga, Mol. Cell. Proteomics 9 (2010) 1133–1143.
36. Y. Kimura, K. Nagata, N. Suzuki, R. Yokoyama, Y. Yamanaka, H. Kitamura, H. Hirano, O. Ohara, Proteomics 10 (2010) 3884–3895.
37. E. Kinoshita, E. Kinoshita-Kikuta, Proteomics (2011) in press.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.50

Large-scale 2-D DIGE studies - guidelines to overcome pitfalls and challenges along the experimental procedure

Franziska Dautel[1], Stefan Kalkhof[1], Saskia Trump[2], Irina Lehmann[2], Andreas Beyer[3], Martin von Bergen*[1,4].

[1]Department of Proteomics, UFZ, Helmholtz-Centre for Environmental Research, Permoserstr. 15, 04318 Leipzig, Germany; [2]Department of Environmental Immunology, UFZ, Helmholtz-Centre for Environmental Research, Permoserstr. 15, 04318 Leipzig, Germany; [3]Biotechnology Center, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany; [4]Department of Metabolomics, UFZ, Helmholtz-Centre for Environmental Research, Permoserstr. 15, 04318 Leipzig, Germany.

Received: 29 October 2010 **Accepted:** 14 December 2010 **Available Online**: 15 December 2010

ABSTRACT

In large 2-D DIGE proteomic studies with a large number of samples, it is essential to design the experimental setup to detect statistically significant protein changes under consideration of experimental variances. Herein are presented guidelines and general remarks on the extraction of protein expression data by following protein spots on their way from first spot synchronization, detection, quantification and statistical analysis until excision and identification. Further discussion addresses common difficulties, potential pitfalls and strategies for dealing with gel-to-gel discrepancies, labeling inefficiencies, and dye- and batch effects which might not be obvious to novices and even more experienced users of DIGE technology.

Keywords: Protein expression analysis, Proteomics; Differential gel electrophoresis, 2-D-DIGE.

Abbreviations.

2-D-DIGE, 2-D-difference gel electrophoresis; **B[a]P,** benzo[a]pyrene.

1. Introduction

A typical task for researches with a variety of biological queries is to detect the up- or downregulation of proteins belonging to two or more biological groups such as treatment versus control. Therefore, it is widely accepted and recommended to conduct more than one biological and at least three technical replicates of each group. A popular technique for protein quantification is the two-dimensional polyacrylamide gel electrophoresis (2D-GE) which allows both the separation and visualization of thousands of protein species and the protein identification and quantification of their expression patterns. The greatest achievement of 2-D differential in gel electrophoresis (DIGE) lies in the simultaneous separation of more than one sample per gel as well as having an internal standard for the relative quantification of spot intensities [1]. Samples are labeled prior to electrophoresis with spectrally resolvable fluorescent cyanine dyes Cy2, Cy3 and Cy5, mixed prior to isoelectric focusing (IEF) and resolved on the same 2-D gel [2]. Although the proteome can be assessed by different means, DIGE has shown to be a sensitive, accurate, reproducible, financially affordable and easy to handle approach. For many laboratories it is the method of choice for a quantitative proteome analysis - especially for the reliable detection of minor changes in protein abundances that are not detectable by other staining methods [3]. The technology allows the processing of large numbers of samples simultaneously, e.g. for measuring dense time courses. However, the data processing is still not fully supported by the available software. In particular, there are no established standard procedures to process the quantitative information extracted from 2-D DIGE experiments involving a large number of gels.

The main focus of this study is to present strategies for

***Corresponding author:** PD Dr. Martin von Bergen. UFZ, Helmholtz-Centre for Environmental Research, Department of Proteomics, Permoserstr. 15, 04318 Leipzig, Germany. Fax: +49-341-2351786. Email Address: Martin.vonbergen@ufz.de.

dealing with gel-to-gel discrepancies, labeling inefficiencies, and dye- and batch effects. Gel-to-gel discrepancies arise from run-time differences, variances in the loaded protein amounts or dye-front deformations [4]. Accounting for these differences is important for both 2-DE and 2-D DIGE. However, the dye-effect is specific for DIGE-projects, as the application of three different fluorophores can cause preferential dye-protein binding, variances in the fluorescent signal and background and differences in gel migration of the labeled proteins. As a result, protein abundances are not directly comparable when the proteins are labeled differently in various samples [5-7]. In addition, the experimental execution for a large number of samples is often divided into several batches of 6 or 12 gels. As a consequence, results of protein expression often cluster with the performed batches rather than with the individual samples and replicates.

The goal is to identify spots that are truly differentially expressed, while accounting for statistical issues such as the multiple testing problem. This multiple testing problem states the accumulation of false positives as a general property of confidence-based statistical tests. These tests are applied across multiple features such as individual spots in DIGE to detect significantly altered changes in protein abundance [8].

This study reports an experimental design for a 2-factor analysis (time and concentration): murine hepatoma cells (Hepa1c1c7) were treated with the procarcinogen benzo-*a*-pyrene (B[a]P) and protein concentrations were quantified using 2-D DIGE (Fig. 1). Differential protein expression induced by B[a]P (or active B[a]P-metabolites) has previously been studied in different cellular models using one incubation time point and several B[a]P- or B[a]P-metabolite concentrations [9-13]. In contrast, this B[a]P-protein expression analysis sampled four incubation time points at one toxic (5 μM) and one sub-acute B[a]P-concentration (50 nM), which required the processing of 36 samples in total [14]. In order to process the data originating from these experiments, a statistical analysis pipeline was developed to account for dye- and batch effects and to extract concentration- and time-dependent protein profiles.

2. Material and Methods

2.1 *Cell culture and BaP exposure*

Murine hepatoma cells (Hepa1c1c7, ATCC No. CRL-2026; LGC Promochem, Wesel, Germany) were cultured as described elsewhere [15]. The cells were exposed to 50 nM B[a]P (Sigma-Aldrich, Steinheim, Germany), 5 μM B[a]P or DMSO for 2, 4, 12 and 24 h. Three independent biological replicates of all treatments were prepared.

2.2 *DIGE and Data Analysis*

2.2.1 Difference gel electrophoresis

Cells were washed and lysed according to the procedure previously described [16]. Protein extracts were prepared and labeled according to manufacturer's recommendations (GE Healthcare, Uppsala, Sweden). A Cy2-labeled common internal standard for all gels was prepared from a mixture of all samples IPG strips (24 cm, pH range 3-10 NL; GE Healthcare, Freiburg, Germany), which were rehydrated overnight and focused for 100,000 Vhrs using an Ettan IPGphor 3 isoelectric focusing unit (GE Healthcare, Freiburg, Germany) as described earlier [17]. Second dimension separation was performed using an Ettan DALTtwelve electrophoresis system (GE Healthcare, Uppsala, Sweden) on 12 % SDS-PAGE gels. The gels were scanned using the Ettan DIGE Imager Scanner (GE Healthcare, Uppsala, Sweden).

2.2.2 *DIGE analysis*

The gel image analysis was performed using Delta 2-D version 3.6 (Decodon GmbH, Greifswald, Germany; [18]). The gels were warped and a fusion gel was created including all gels of the experiment. Subsequent to the spot detection, the spots were manually edited and transferred to all individual gel pictures. Relative spot volumes (integrated staining intensities) were determined by normalizing the spot volumes to the total protein amount on each gel (excluding the largest spots representing ~ 5 % of the total intensity). The relative spot volumes were extracted and transformed to a log2-scale. Afterwards the log2-ratios of the Cy3 and Cy5 intensities were adjusted to the intensities of the internal standard on the Cy2 channel. A dye-specific bias was observed. In order to remove the bias, the effect of the dye type (Cy3 or Cy5) was regressed out and the residuals were used for subsequent analysis. The distribution of residual spot intensities on each gel was centered by subtracting the mean of each gel. A random effects model was fitted to each spot to account for the spot-specific intra-gel correlation between the Cy3 and Cy5 signals. To eliminate the batch effect, the mean of the DMSO samples at the respective time point were subtracted from the residuals of this model. Thus, each final measurement should only reflect perturbation due to B[a]P exposure.

A two-way ANOVA model with B[a]P-exposure time and concentration as the factors was fitted to each spot. P-values for the time main effect, the concentration main effect, and their interaction were corrected for multiple hypotheses testing using the false discovery rate (FDR). The time main effect was significant for 120 spots at FDR < 0.05. Only these spots were considered in the subsequent analysis.

2.2.3 *Preparation of 2-D-reference gels for protein identification*

Since DIGE-gels only contain 300 μg protein/gel, only very large protein spots are visible on DIGE-gels after applying the blue silver staining method - a modified Neuhoff's colloidal Coomassie Blue G-250 staining with sensitivity close to silver staining [19] (data not shown). In order to detect all identified protein spots, reference gels with 2.0 and 2.5 mg protein (equal mix of all samples) were created for protein identification (procedure as described elsewhere [16]). To avoid mistakes in protein spot identification, the images of the

reference gels were loaded in the Delta 2-D DIGE-project and warped to the DIGE-images.

2.3 *Protein Identification by MALDI-MS or nano-HPLC/ESI-MS*

Following spot excision from the reference gel and tryptic digestion, the measurement was performed with MALDI-TOF/TOF-MS (Ultraflex III, Bruker Daltonik, Bremen Daltonik, Bremen, Germany) using the HCCA matrix (0.6 mg/ml) according to Georgieva *et al.* [20]. Alternatively, if no significant identification was obtained with MALDI-TOF-MS analysis (Mascot-Score cut-off 100) the samples were measured using a nano-HPLC system (2-D-nano-HPLC, Eksigent, Dublin, CA, USA) coupled to an LTQ-Orbitrap XL ETD hybrid mass spectrometer (Thermo Fisher Scientific, USA) [21].

3. Results and Discussion

3.1 General workflow

The analysis of murine hepatoma cells (Hepa1c1c7) treated with the procarcinogen B[a]P is presented as an example showing how to conduct and analyze large 2-D DIGE studies. In this project both different B[a]P-concentrations and incubation time points are used (Table 1). In the optimized workflow (Fig. 1) Cy2 normalization (section 4.3.1) as well as corrections of the labeling (4.3.2) and gel batch effect (4.3.3) were incorporated before the final expression patterns were extracted and tested for significant regulation by ANOVA and FDR estimation. Spots of interest with significant expression changes were identified using post-stained reference gels which were warped to the equivalent DIGE-gels in the Delta2D-project to eliminate false identifications. Each step is discussed in detail in the following sections.

3.2 *Spot matching and spot detection using Delta2D*

The first and crucial step in the 2-D DIGE workflow before data processing involves the recognition of common spots across different gels (Fig. 1). In the proposed setup, this part is performed using the Delta2D software. However, free, open-source image-processing algorithms for image registration and fusion are also available [22]. The time consuming image analysis is one of the bottlenecks in 2-DE studies and the assistance of a manual operator is needed [23]. Thus, possible drawbacks of this workflow-step have an important impact on the outcome of the further analysis.

Twenty 2-D DIGE gels (24 cm, p*I* 3-10) were created to analyze two B[a]P concentrations, DMSO as a control and four different time points in triplicates (36 samples) (Table 1). All gel images were loaded into the Delta2D software. However, uploading and processing such a large number of images can lead to severe software instabilities. In this case, only the 64 bit version of Delta2D (not the 32 bit version) was able to administer 60 gel images in one project at once.

Another problem is the visualization of large gel series: it is impossible to show all gels side-by-side, which can lead to confusion and consequently mistakes in the analysis by the manual operator [18].

Table 1. Experimental setup for the large 2-D DIGE study of B[a]P-induced alterations in protein expression in murine Hepa1c1c7 cells. Gel batches were performed according to incubation time points and dyes were swapped among biological replicates and among the DMSO-samples of the 3rd replicate in each time point.

Sample	Time	Replicate	Label	Gel No.	Batch
5 µM B[a]P	2 h	1	Cy3	2	1
50 nM B[a]P	2 h	1	Cy5	1	1
DMSO	2 h	1	Cy3	1	1
5 µM B[a]P	2 h	2	Cy5	3	1
50 nM B[a]P	2 h	2	Cy3	3	1
DMSO	2 h	2	Cy5	2	1
5 µM B[a]P	2 h	3	Cy3	5	1
50 nM B[a]P	2 h	3	Cy5	4	1
DMSO	2 h	3	Cy3/Cy5	4+5	1
5 µM B[a]P	4 h	1	Cy5	7	2
50 nM B[a]P	4 h	1	Cy3	6	2
DMSO	4 h	1	Cy5	6	2
5 µM B[a]P	4 h	2	Cy5	8	2
50 nM B[a]P	4 h	2	Cy3	8	2
DMSO	4 h	2	Cy3	7	2
5 µM B[a]P	4 h	3	Cy3	9	2
50 nM B[a]P	4 h	3	Cy5	10	2
DMSO	4 h	3	Cy3	9+10	2
5 µM B[a]P	12 h	1	Cy3	12	3
50 nM B[a]P	12 h	1	Cy5	11	3
DMSO	12 h	1	Cy3	11	3
5 µM B[a]P	12 h	2	Cy5	13	3
50 nM B[a]P	12 h	2	Cy3	13	3
DMSO	12 h	2	Cy5	12	3
5 µM B[a]P	12 h	3	Cy3	15	3
50 nM B[a]P	12 h	3	Cy5	14	3
DMSO	12 h	3	Cy3/Cy5	14+15	3
5 µM B[a]P	24 h	1	Cy3	17	4
50 nM B[a]P	24 h	1	Cy5	16	4
DMSO	24 h	1	Cy3	16	4
5 µM B[a]P	24 h	2	Cy5	18	4
50 nM B[a]P	24 h	2	Cy3	18	4
DMSO	24 h	2	Cy5	17	4
5 µM B[a]P	24 h	3	Cy3	20	4
50 nM B[a]P	24 h	3	Cy5	19	4
DMSO	24 h	3	Cy3/Cy5	19+20	4

2-DE-gels are often affected by spatial distortions due to run-time differences and dye-front deformations. Spot matching is the first and essential step to receive a proteome map of good quality. This is achieved via 'warping' the images in the image analysis software to remove distortions from the gel images and to bring the spot patterns into congruency [18]. Delta2D has been shown to be a fast and more reliable image analysis software than comparable commercial products like Proteomweaver (Definiens), especially due to its precise warping procedure [24].

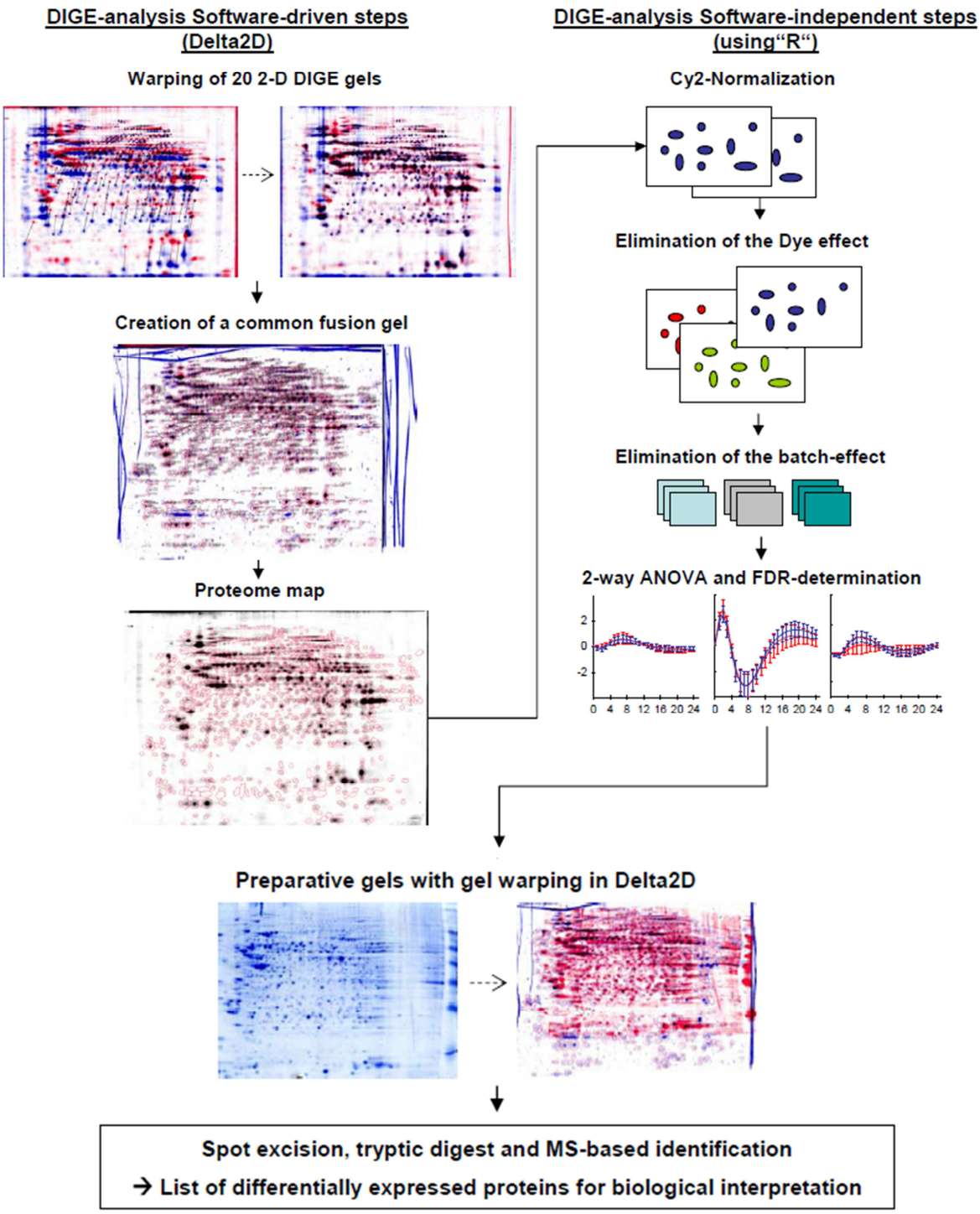


Figure 1. *Flowchart to demonstrate the individual steps in the proposed 2D-DIGE analysis.* Gel images are loaded in appropriate analysis software (e.g. Delta2D) for gel warping, creation of a fusion gel and the subsequent spot detection to create a common proteome map, which is then transferred on all gels. To correct for gel- to gel-variation, dye- and batch effects and to extract time- and concentration dependent protein expression curves, it is recommended to transfer the data to independent statistical software such as "R".

A key advantage of using a pooled internal standard is that the same sample is used for the gel-to-gel matching, as opposed to matching gels that contain different samples and that may consequently have different spot patterns [2]. The alignment of gel images establishes a spot consensus pattern by creating a composite image summarizing the whole experiment's gel information (Fig. 1). A robust reference image is particularly important for large studies since the large number of gels processed in different batches increases the likelihood of detecting artifactual differences simply by

chance. Manual inspection of this virtual fusion gel verified that it did not contain visible imperfections such as dust, air bubbles or precipitated dye. In order to be declared as a spot, a dot on the fusion gel was required to match the three-dimensional profile characteristics of a spot. With the consensus spot patterns transferred to all gels from the fusion gel, 1227 spots were detected. In comparison, Corzett *et al.* applied a more traditional approach (DeCyder Differential Analysis Software v5.01, GE Healthcare) using spot detection on each individual gel resulting in different spot patterns for each gel. Thus, only 165 (6.8 %) of all detected spots could be matched on all 12 gels loaded with protein lysates of human plasma [5]. In conclusion, the detection of differentially expressed proteins can be substantially improved by using consensus spot patterns [24]. Without a unified proteome map and thus separate spot detections on every gel, the missing values must be engrafted by statistical means such as missing value imputation [18].

However, one pitfall in the 2-D DIGE setup remains: the need for a manual operator in spite of the associated great dependency on the performance of the individual. The decision whether a spot is a spot and where it begins or ends can be a source for unintentional data manipulation. Although automated warping procedures exist, they do not yet provide sufficient quality, especially for a large number of gels in which artifacts accumulate and gel-to-gel variations increase due to different gel batches. It has been observed that vectors were set incorrectly, requiring time-consuming manual corrections [23, 24]. In addition, automated spot matching is prone to spot amalgamation in regions of variable spot resolution and can lead to erroneous measurements [25]. Thus, a future goal should be the reduction of user controlled settings, as described in [26].

3.3 *Data analysis and normalization*

3.3.1 *Data processing and Cy2-Normalization*

Existing software analysis packages specialized for 2-D DIGE evaluation lack important features required for analyzing larger sets of gels. Large 2-D DIGE datasets with various time points and concentrations are in need of a more flexible analysis than datasets with only two samples (e.g. treated versus control) in three replicates.

The following steps have been taken to deal with problems specific to large 2-D DIGE projects (Fig. 1). First, a pooled internal standard labeled with Cy2 was included in the 2-D DIGE experimental design. Second, using 3 different fluorophores for pre-protein staining can result in a dye-specific effect. In Delta2D it is possible to detect a dye-effect if a dye-swapping was performed (refer to section 4.3.2), but the data has to be transferred to external software to correct for this effect (Fig. 1). Third, a gel-specific batch effect was observed for the data which was confounded with the B[a]P-incubation time points (section 4.3.3). Again, Delta2D only allows for detecting but not correcting batch effects. The ultimate goal was to extract concentration- and time-dependent protein expression curves. Delta2D can extract either concentration- or time-dependent expression profiles; however, a combined graph can not be exported.

In conclusion, the protein spot volume data were exported and all subsequent analysis was done using external statistics software, R in this case (www.r-project.org). Delta2D reports spot volumes as percent of the total volume on the respective gel. These percentages were transformed into log2-fractions, which is more convenient since this scale is symmetrical and centered at zero. For example, a two fold increase and decrease on the log2 scale is 1 and -1, respectively, whereas on the percentage scale the same changes would be 200 % and 50 %, respectively. This lack of symmetry can pose problems for downstream-analysis. In addition, spots with missing data were removed. Next, the ratio of the Cy3 and Cy5 channels to the internal standard (Cy2) for each spot was calculated and transformed to the log2 scale, eliminating false effects arising from slightly different protein amounts loaded on the gels. In addition to having the same spot patterns for spot matching to receive a common proteome map, this is the second advantage of the internal standard approach. However, one possible problem of using the Cy5/Cy2 and Cy3/Cy2-ratios (as also suggested by the Delta2D software) is the violation of the statistical assumption of independent sampling [4]. In addition, spatial trends in the intensity measurements of the internal standard were observed and one-third of the available sample-space is lost [25]. Engelen *et al.* suggested removing the spatial bias with a strategy called spatial intensity bias removal (SIBR) using a two-dimensional nonlinear regression algorithm. Since SIBR does not need an internal standard, more biological samples could be run on a single gel, resulting in decreasing numbers of gels, lower costs, shorter sample preparation and image analysis steps [25]. Another approach suggests one sample (test or control) with a pooled standard on one gel as performed in saturation labeling also in minimal labeling approaches [27]. But with limited resources, restricted numbers of gels and many samples to be analyzed, the three-dye minimal approach is more readily applicable [28]. Furthermore, the use of an internal standard remains the most applicable approach to correct for gel-to gel variations.

3.3.2 *Correcting the dye effect*

One systematic source of variation within the dataset could come from a dye effect (Fig. 1). Often it is assumed that a protein-specific dye effect does not occur. As a result, protein abundances should be directly comparable across dyes after normalizing the Cy3 and Cy5-values to the internal standard Cy2 as suggested by Delta2D [18]. However, several studies have demonstrated an existing dye effect in DIGE-studies due to a combination of preferential dye binding to proteins, differences in gel migration of the labeled proteins, and differences in the fluorescent signal and background [5-7]. One method of eliminating dye effects is to perform a dye-swap. In the B[a]P-DIGE dataset, the Cy3/Cy5-fluorophores were switched among the samples (Table 1) ensuring that

within the three biological replicates for each sample, dyes were swapped at least once. The dye swap would generally double the number of gels needed as each sample would have to be measured twice, or only half of the biological replicates would be left in the experiment. A sensible solution that avoids such a suboptimal experimental design is to swap the two dyes among the three biological replicates, one dye being the same in the corresponding sample of two replicates. Additionally, one sample per batch (usually the DMSO-control of the 3rd replicate) was labeled with both dyes on separate gels (Tab. 1). Thus, a distinction can be made between the dye effect and differences coming from the biological replicates. With this approach, a protein-specific dye effect could clearly be detected. Such protein-specific biases require computational correction, especially if a large number of samples are analyzed and not all samples can be swapped individually [5]. Within Delta2D, a mathematical dye correction cannot be performed. Corzett *et al.* described a method to account for spot variances, including the dye effect, based on a mixed-effects model [5]. In this case, the dye correction was done by regressing out the effect of the dyes and taking the residuals using a fixed effects model. Following this, distribution of log2 ratios for each gel were centered. In the second step, a random effects model was fitted to account for the correlation between samples which were run on the same gel.

3.3.3 *Correcting the batch effect*

In customized GE systems, it is often only possible to run a small number of gels in one batch. Eravci *et al.* extended the common available GE-apparatus to the load of 24 gels at once [24]. However, the handling of up to twelve gels at once in the commercially available instruments provides a wide variability of possible errors in the sample and gel processing, resulting in false results or artifacts and avoidable repetitions of experiments. In addition, technical failures or mistakes along the experimental procedure can lead to gel artifacts rendering entire batches useless for any further analysis. The costs for losing complete batches of 12 2-D DIGE-gels or more must be considered.

For this study, the twenty gels were divided in batches according to B[a]P-incubation time points (Tab. 1, Suppl. Fig. 1). Even after the correction of all spot intensities of Cy3- and Cy5-labelled samples with the common internal standard Cy2, the batch effect was still visible (Suppl. Fig. 1). To differentiate between the time- and the batch effect, the data of the control samples DMSO were used. Based on the results from cytotoxicity measurements and considering that the cells are not synchronized, it was concluded that any temporal effects observed in the control samples are not due to different incubation times, but artifacts caused by the batch effect [14]. Thus, the DMSO mean from each incubation time point was subtracted from the corresponding B[a]P-incubated samples. The batch effect correction leads to an increased confidence that the observed change in protein abundance is due to biological happening in the cell and not due to experimental variability (Fig. 2B). The batch effect increases with the number of samples as more runs are needed to process all samples. Furthermore, it has been demonstrated that commercially available IPG strips show large differences when they originate from different mastergels. Often the strip numbers in one package are not consecutive (GE Healthcare) or not numbered at all (other suppliers). Consequently the strips originate from different mastergels. This contributes to the batch effect as well [24].

3.3.4 *Identification of significantly changing protein spots*

After completion of the corrections outlined above, the next step is determining proteins exhibiting significant changes as functions of time or concentration. The statistical analysis of DIGE software packages like DeCyder (GE Healthcare), Progenesis SameSpots V3.0 (Nonlinear Dynamics) and Dymension 3 (Syngene) (comparable to Delta2D) have been shown to be inconsistent; results obtained for protein fold changes for one dataset were substantially different in each package. Thus, DIGE quantification is still software dependent despite the use of an internal standard [26]. This observation calls for a reevaluation and validation of the results by DIGE analysis software independent statistical methods such as the external statistics software "R", and biological methods, such as Western Blot.

Many studies exclusively focus on the calculation of protein abundance ratios, often expressed as fold changes (e.g. control versus treatment). Additionally, the student's t-test is often used to test for statistical significance of concentration changes. Such univariate methods determine whether the differences between two samples are significant. However, such methods are inappropriate for studies involving more than two conditions (i.e. more than one treatment condition). In this case two factors exist, time and concentration, time having four levels (4 time points). The t-test in such a situation would not discern whether differences between samples are caused by time or concentration effects, because most samples differ with respect to both time and concentration. Shen *et al.* reported protein changes in human amniotic epithelial FL cells incubated with 0.05 μM anti-benzo[*a*]pyrene-7,8-dihydrodiol-9,10-epoxide in which no single protein was significantly altered in all three incubation time points [12].

Therefore study designs involving multiple factors require the application of Analysis of Variance (ANOVA). ANOVA determines the extent to which the observed variances between samples can be explained by the experimental parameters as opposed to biological or technical variation within the experiment [6]. Delta2D provides a large variety of statistical tools including a two-way ANOVA model. However, to analyze the variances among this data and account for dye and batch effect, the raw data must be extracted from Delta2D for further processing, precluding any possible return to the gel analysis software for that project. Thus, within the two-way ANOVA model, the concentration factor (5 μM and 50 nM B[a]P) and the time factor (2 h, 4 h, 12 h, 24 h) as well as their interaction were considered. This model was fit for

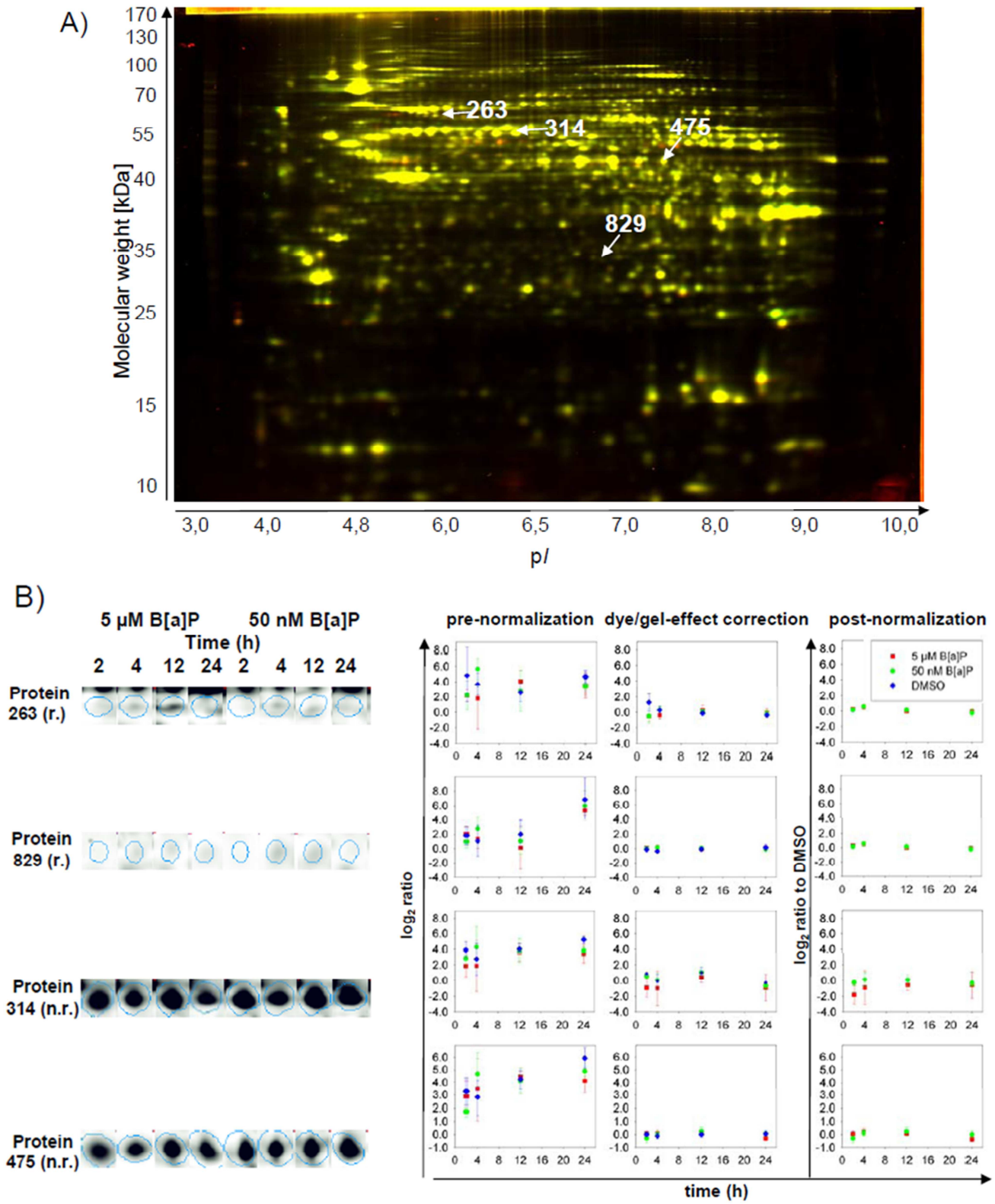


Figure 2. *Visualization of protein expression results. A)* 2-D DIGE gel of murine Hepa1c1c7cell-protein extract. The cells were incubated for 4 h with 50 nM B[a]P(Cy3) and DMSO (Cy5). *B)* Extracted protein spots (left) in comparison with the time- and concentration-dependent protein expression profile (right) of 4 selected proteins (263, 829, 314, 475). Whereas a spot album gives rather visual information about the protein abundance without any normalization, the expression curves show the entire behavior of a protein over B[a]P-concentration (5 µM: red; 50 nM: green; DMSO: blue) and exposure time based on normalized data. Displayed are the log2-ratios of the spot intensities after taking the Cy3/Cy2 and Cy5/Cy2 ratios (pre-normalization), following the dye/gel-correction and after correcting for the batch effect (post-normalization). Proteins 263 and 829 were differently regulated over B[a]P-exposure time (r.) while the expression of proteins 314 and 475 were not found to be changed (n.r.). The location of these spots on a 2-D DIGE gel is also marked in A).

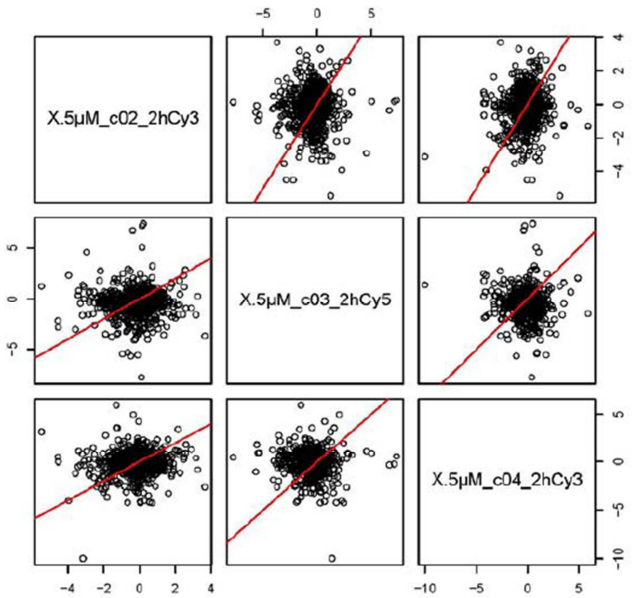
each spot and the p-values for each effect were recorded. Finally, an adjustment for multiple testing was made. Each test has a certain probability of giving false positive results. In the matter of 2-D DIGE, a protein spot could be declared significantly altered in expression, despite the difference being due only to chance. This problem is exacerbated for experiments involving a large number of tests, as is commonly done when using DNA microarrays. Due to the large number of gels processed in this study, similar problems were faced, and inflated false positive rates were accounted for by computing the False Discovery Rate (FDR). The FDR is the expected rate of false positives among all results that were declared positive (i.e. 'significant') [4, 29]. After the FDR adjustment, 120 out of 1227 protein spots had a significant time effect at FDR < 0.05 and no spots exhibited a significant concentration effect or an interaction between time and concentration (Fig. 2A).

3.4 *Protein spot excision*

A crucial step for obtaining reliable results in 2-D DIGE experiments is the excision of significantly altered protein spots from gels (Fig. 1). Spot excision can be performed manually or by using spot pickers. To increase the detection rate of differentially expressed proteins of 2-D DIGE analysis without the availability of a spot picker and still eliminating false spot identifications, a two-step approach is proposed: after staining the DIGE-gels with Coomassie Brilliant Blue (CBB) or equivalent stains, the images should be loaded in the analysis software and warped within the 2-D DIGE project to the internal standard of the gel. This ensures that the same spot, labeled once with Cy3, Cy2 or Cy5 and once with a post-stain, is being detected. In most cases, approximately 100 µg of protein are labeled for each sample and thus 300 µg protein lysate in total are loaded on one 2-D DIGE gel. If the separation is increased by using large IEF-strips and long gels for the second dimension, low abundant spots are often not detectable with post-staining methods (Suppl. Fig. 2). Therefore, in the second step, 2-D reference gels are prepared carrying a larger protein amount (in this case 2.0 and 2.5 mg) from a mixture of all samples incubated with the same post-stain, and are additionally warped to the fusion gel of the 2-D DIGE project to ensure that exactly the same spots are excised. Ideally, one would use one reference gel for each type of sample (e.g. treated versus control), as the spot pattern might then be similar to the respective DIGE-gel. Each reference gel would then be warped to the respective samples e.g. using Delta2D. However, such an approach is often impossible due to limited resources.

Post-staining with other dyes prior to spot picking has been recommended even when using spot pickers [30-32]. Slight mobility differences between labeled and unlabeled species of the same protein have been reported. The unlabeled protein moves slightly faster than the labeled equivalent (about 0.5 kDa) [2] which can cause problems in the subsequent MS-based identification of the protein. However, proteins may be missed, since different proteins vary in their individual staining properties and some post-stains show relatively low sensitivity such as CBB [30]. About 40% of the differentially regulated spots discovered with DIGE could not be reliably detected after post-staining with colloidal CBB [32]. Better results were reported for SYPRO Ruby, but its excitation is achieved either with UV light or with laser scanners. Thus,

A) Normalization achivied with Delta2D



B) Dye/Gel Normalization

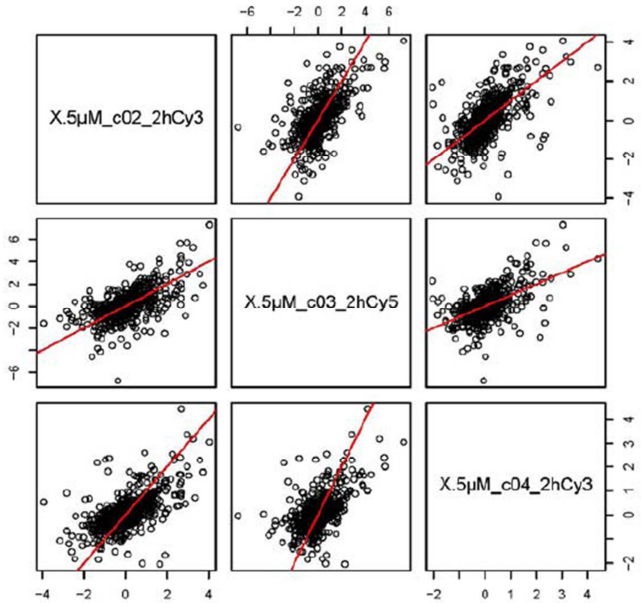


Figure 3. *Comparison between the different normalization approaches (Delta2D-based and -independent). A)* With no correction for the different variances in the gels (dye- and batch effect), the sample replicates do not show any linear trend. *B)* Note the improvement in the linear correlation between the sample replicates after correcting for gel-, dye- and batch effects in this approach.

spots would have to be picked under excitation and not under daylight condition [31]. In this study, the "blue silver" staining was used, which is faster and more sensitive in comparison to colloidal CBB, and spots can be cut in daylight under a hood to prevent ceratin contamination [19]. In the end, the excised spots were digested and identified using MALDI- or ESI-MS/(MS), as it is often useful to combine different types of MS [14, 28].

4. Concluding remarks

Evaluation of proteomic data with 2-D DIGE technology remains a multistep process. In this study, the route of the protein spots is followed along the experimental procedure. From the appropriate gel warping and spot detection to the final spot excision, the researcher faces multiple challenges for obtaining high-quality protein expression data. It has been shown that protein quantification can be improved by taking into account dye and batch effects and proper statistical analysis (Fig. 3). Some difficulties concerning 2-D DIGE experimentation and evaluation must still be overcome, particularly the large individual influence of the manual operator, robust correction for different variances during the spot analysis, and correct spot excision call for improved methods. Reliable protein expression time courses for B[a]P treated murine Hepa cells were successfully extracted (Fig. 2B). These expression curves will provide further hints on the molecular processes triggered by B[a]P-exposure.

5. Supplementary material

Supplementary material regarding this manuscript is online available in the web page of JIOMICS.

http://www.jiomics.com/index.php/jio/rt/suppFiles/50/0

Acknowledgements

The authors thank Jacob Michaelson, Dr. Janina Tomm, Stephan Müller, David Lee Harris, Yvonne Kullnick, Antje Thonig, Anna Lohse and Kerstin Krist for cooperation and technical assistance. The project was supported by the Helmholtz-Alliance on Systems Biology and by Helmholtz Impulse and Networking Fund through Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE). AB acknowledges funding from the Klaus Tschira Foundation.

References

1. Unlu, M., M.E. Morgan, and J.S. Minden, Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. Electrophoresis, 1997. **18**(11): p. 2071-7.
2. Alban, A., et al., A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. Proteomics, 2003. **3**(1): p. 36-44.
3. Karp, N.A., et al., Comparison of DIGE and post-stained gel electrophoresis with both traditional and SameSpots analysis for quantitative proteomics. Proteomics, 2008. **8**(5): p. 948-60.
4. Karp, N.A., et al., Experimental and statistical considerations to avoid false conclusions in proteomics studies using differential in-gel electrophoresis. Mol Cell Proteomics, 2007. **6**(8): p. 1354-64.
5. Corzett, T.H., et al., Statistical analysis of the experimental variation in the proteomic characterization of human plasma by two-dimensional difference gel electrophoresis. J Proteome Res, 2006. **5**(10): p. 2611-9.
6. Karp, N.A. and K.S. Lilley, Maximising sensitivity for detecting changes in protein expression: experimental design using minimal CyDyes. Proteomics, 2005. **5**(12): p. 3105-15.
7. Krogh, M., et al., Analysis of DIGE data using a linear mixed model allowing for protein-specific dye effects. Proteomics, 2007. **7**(23): p. 4235-44.
8. Listgarten, J. and A. Emili, Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. Mol Cell Proteomics, 2005. **4**(4): p. 419-34.
9. Gao, Z., et al., Zinc finger proteins and other transcription regulators as response proteins in benzo[a]pyrene exposed cells. Mutat Res, 2004. **550**(1-2): p. 11-24.
10. Hooven, L.A. and W.M. Baird, Proteomic analysis of MCF-7 cells treated with benzo[a]pyrene, dibenzo[a,l]pyrene, coal tar extract, and diesel exhaust extract. Toxicology, 2008. **249**(1): p. 1-10.
11. Lau, A.T. and J.F. Chiu, Proteomic and biochemical analyses of in vitro carcinogen-induced lung cell transformation: synergism between arsenic and benzo[a]pyrene. Proteomics, 2006. **6**(5): p. 1619-30.
12. Shen, W., H. Liu, and Y. Yu, Translation initiation proteins, ubiquitin-proteasome system related proteins, and 14-3-3 proteins as response proteins in FL cells exposed to anti-benzo[a]pyrene-7,8-dihydrodiol-9,10-epoxide. Proteomics, 2008. **8**(17): p. 3450-68.
13. Shen, W., H. Liu, and Y. Yu, Proteomic analysis of cellular responses to different concentrations of anti-benzo(a)pyrene-7,8-dihydrodiol-9,10-epoxide in human amniotic epithelial FL cells. J Proteome Res, 2007. **6**(12): p. 4737-48.
14. Dautel , F., Kalkhof , S., Trump , S., Michaelson , J., Beyer , A., Lehmann , I., von Bergen, M., DIGE based protein expression analysis of BaP-exposed hepatoma cells reveals a complex stress response including alterations in oxidative stress, cell cycle control, and cytoskeleton motility at toxic and subacute concentrations. J. Proteome Res., 2010.
15. Solhaug, A., M. Refsnes, and J.A. Holme, Role of cell signalling involved in induction of apoptosis by benzo[a]pyrene and cyclopenta[c,d]pyrene in Hepa1c1c7 cells. J Cell Biochem, 2004. **93**(6): p. 1143-54.
16. Morbt, N., et al., Proteome changes in human bronchoalveolar cells following styrene exposure indicate involvement of oxidative stress in the molecular-response mechanism. Proteomics, 2009. **9**(21): p. 4920-33.
17. Benndorf, D., et al., Identification of spore allergens from the indoor mould Aspergillus versicolor. Allergy, 2008. **63**(4): p. 454-60.
18. Berth, M., et al., The state of the art in the analysis of two-dimensional gel electrophoresis images. Appl Microbiol Biotechnol, 2007. **76**(6): p. 1223-43.
19. Candiano, G., et al., Blue silver: a very sensitive colloidal Coomassie G-250 staining for proteome analysis. Electrophoresis, 2004. **25**(9): p. 1327-33.
20. Georgieva, D., et al., Comparative analysis of the venom proteomes of Vipera ammodytes ammodytes and Vipera ammodytes meridionalis. J Proteome Res, 2008. **7**(3): p. 866-86.
21. Muller, S.A., et al., Optimization of parameters for coverage of

low molecular weight proteins. Anal Bioanal Chem.

22. Sorzano, C.O., et al., Elastic image registration of 2-D gels for differential and repeatability studies. Proteomics, 2008. **8**(1): p. 62-5.
23. Millioni, R., et al., Delta2D and Proteomweaver: Performance evaluation of two different approaches for 2-DE analysis. Electrophoresis. **31**(8): p. 1311-7.
24. Eravci, M., et al., Improved comparative proteome analysis based on two-dimensional gel electrophoresis. Proteomics, 2007. **7**(4): p. 513-23.
25. Engelen, K., et al., Alternative experimental design with an applied normalization scheme can improve statistical power in 2D-DIGE experiments. J Proteome Res. **9**(10): p. 4919-26.
26. Kang, Y., et al., Comparison of three commercially available DIGE analysis software packages: minimal user intervention in gel-based proteomics. J Proteome Res, 2009. **8**(2): p. 1077-84.
27. Karp, N.A. and K.S. Lilley, *Design and analysis issues in quantitative proteomics studies.* Proteomics, 2007. **7 Suppl 1**: p. 42-50.
28. McNamara, L.E., et al., Fluorescence two-dimensional difference gel electrophoresis for biomaterial applications. J R Soc Interface. **7 Suppl 1**: p. S107-18.
29. Benjamini, Y., et al., *Controlling the false discovery rate in behavior genetics research.* Behav Brain Res, 2001. **125**(1-2): p. 279-84.
30. Miller, I., J. Crawford, and E. Gianazza, *Protein stains for proteomic applications: which, when, why?* Proteomics, 2006. **6**(20): p. 5385-408.
31. Gharbi, S., et al., Evaluation of two-dimensional differential gel electrophoresis for proteomic expression analysis of a model breast cancer cell system. Mol Cell Proteomics, 2002. **1**(2): p. 91-8.
32. Tonge, R., et al., Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. Proteomics, 2001. **1**(3): p. 377-96.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v1i1.29

Transcriptomic responses in Japanese medaka (*Oryzias latipes*) exposed to 17β-estradiol

Mizukami-Murata Satomi[1], Kishi-Kadota Katsuyuki*[1].

[1]Japan Pulp & Paper Research Institute, Inc., Tokodai 5-13-11, Tsukuba, Ibaraki 300-2635, Japan.

Received: 7 August 2010 **Accepted:** 28 January 2011 **Available Online:** 1 February 2011

ABSTRACT

The effects of 17β-estradiol (E2) were evaluated using the medaka DNA microarray representing 36,398 genes. We first evaluated chronic effects on medaka exposed to E2 at different concentrations for 60 days posthatch. At ≥ 30 ng/L of E2 severe reproductive impairments such as sex reversal were observed. Larval medaka, *Oryzias latipes*, (within 24 hrs posthatch) were then exposed to E2 at various concentrations (3, 30, 100 ng/L) for up to 7 days. Microarray analyses of the E2-exposed larvae revealed that exposure to E2 up-regulated and down-regulated 339 and 105 genes, respectively. The up-regulated genes included ones involved in the p53 signaling pathway, apoptosis, and growth and development, in addition to well-known biomarkers such as vitellogenin and choriogenins. Down-regulated genes included heat shock proteins and estrogen receptors. Most of the up-regulated genes encoding the p53 signaling pathway, apoptosis, and growth and development exhibited a dose-dependent increase in gene expression, whereas the down-regulated genes in the heat shock protein category showed a dose-dependent decrease in gene expression. Time course experiments suggested that the E2 treatment attenuated the time-dependent changes in gene expressions of these genes. Among the genes related to oocyte maturation, estrogen-regulated genes such as choriogenins and vitellogenins were dramatically induced in response to E2 exposure, whereas other steroid-regulated genes such as zona pellucida-domain proteins did not change in gene expression by the E2 treatment. Results suggest that transcriptomic studies on larval medaka help elucidate the effects caused by endocrine disruptors on various biological pathways in vertebrate development.

Keywords: Affinity chromatography; 17β-estradiol; medaka; DNA microarray; Endocrine disruption; Feminization.

1. Introduction

There is a weight of evidence that humans and wildlife are susceptible to endocrine disruption by various natural and synthetic chemicals introduced into the environment [1]. Endocrine disrupting chemicals (EDCs) have the potential to constitute a threat to the reproductive health of organisms, increase the risk of tumor development, and may cause other adverse effects on differentiation, growth, and development [1]. During the past decade, there have been numerous attempts to elucidate the mechanism of endocrine disruption in vertebrates, particularly in fish. Teleost fish, such as Japanese medaka (*Oryzias laptipes*) [2-5], zebrafish (*Danio rerio*) [6, 7], and fathead minnow (*Pimephales promelas*) [8], have been models for mechanistic studies on endocrine disruption. Studies have revealed that many EDCs alter endocrine function by interacting hormone receptors, such as estrogen receptors (ERs), androgen receptors (ARs) and thyroid hormone receptors [9]. Recent studies have shown that reproduction in vertebrates can be disrupted by modulating the hypothalamic-pituitary-gonadal (HPG) axis, by either directly interacting with sex hormone receptors, or by compensatory actions [10]. Various molecular biomarkers have been suggested for screening and testing EDCs, including the yolk protein precursor vitellogenin (VTG) [11], nuclear hormone receptors (e.g., ER and AR) [12, 13], steroidogenic enzymes (e.g., aromatase) [14], and gonadotropins [15]. It is, however, difficult to interpret adverse effects on complicated biological processes based on changes in expression of a few genes. There is clearly a demand for more comprehensive and integrated approaches to better understand the physiological status of an organism affected by EDCs.

***Corresponding author:** Kishi-Kadota Katsuyuki. Postal address: Japan Pulp & Paper Research Institute, Inc., Tokodai 5-13-11, Tsukuba, Ibaraki 300-2635, Japan; Tel.: +81 29 847 9815; Fax: +81 29 847 8923; Email Address: kadota@jpri.co.jp.

Microarray technology provides the mRNA expression profile of virtually all known genes simultaneously to gain a more comprehensive assessment of the chemical impacts on organisms [16]. A transcriptomic approach has the potential not only to allow an insight into mechanisms of toxicity, but also to provide informative data on novel toxic modes of action. Consequently, transcriptomic studies help link unique gene expression profiles elicited by toxicant exposures and biological pathways. For example, gene expression profiling of EDC-exposed fish may aid in the elucidation of effects caused by the EDCs on various biological pathways besides the HPG axis, including metabolism, biosynthesis, secretion, energy production and cell growth, resulting in a more comprehensive understanding of endocrine disruption.

In this work, we investigated the effects of 17β-estradiol (E2) exposure on larval medaka using a medaka microarray representing 36,398 genes. Transcriptional responses of genes were examined to elucidate the effects of the exogenous E2 exposure on various biological pathways in developmental stages of medaka. We found that genes involved in the p53 signaling pathway, apoptosis, growth and development were up-regulated by the exogenous E2, in addition to well-known estrogen-dependent biomarkers such as VTGs and choriogenins (CHGs). By contrast, genes such as heat shock proteins (HSP) and ERs were down-regulated by E2 treatment. Time course experiments suggest that E2 treatment appeared to attenuate the time-dependent changes of some gene expressions in developmental stages. Results indicate that transcriptomic approaches can be used for integrated evaluation of endocrine disruption of developmental stages.

2. Materials and methods

2.1 *Test chemicals*

17β-Estradiol (E2) was obtained from Sigma Chemical Industries, Ltd. (St. Louis, MO, USA) and dimethyl sulfoxide was obtained from Wako Pure Chemical Industries, Ltd. (Osaka, Japan). An E2 stock solution was prepared by dissolving E2 in dimethyl sulfoxide.

2.2 *Test organism*

Japanese medaka (*Oryzias laptipes*, orange-red strain) were originally obtained from the National Institute for Environmental Studies (Tsukuba, Japan), and have been maintained at the Japan Pulp & Paper Research Institute, Inc. The brood stock was maintained at 24 ± 1°C in UV-disinfected, dechlorinated, carbon-filtered tap water with a 16 h light-8h dark photoperiod. The fish were fed *Artemia nauplii* (<24 h after hatching) twice a day.

2.3 *Exposure design*

Exposures of E2 were carried out by a continuous flow-through system, designed to maintain constant concentrations of E2 throughout the exposure experiments. Briefly, the E2 stock solution was delivered to a mixing glass chamber by a mini-chemical pump unit (Oriental Mortar Co., Ltd., Tokyo, Japan) and diluted with carbon-filtered tap water continuously delivered at a flow rate of 100 mL/min. The E2 test solution overflowed the mixing chamber to enter each test glass aquaria containing 23L of the test solution.

For chronic experiments, 25 larval medaka (within 12 hours posthatch) were exposed to 0 (control), 1, 3, 10, 30, and 100 ng/L of E2 (dimethyl sulfoxide with a final concentration of 1:2,500 v/v water) until 60 days posthatch. At the end of the exposures, phenotypic males and females were selected and separated (as determined by secondary sexual characteristics of the fins) for further reproduction tests.

For microarray experiments, larvae were exposed to E2 for 1, 2 and 7 days by the continuous flow-through system as for the chronic exposures except in 1L test chambers. E2 concentrations used were 0 (control), 3, 30, and 100 ng/L for 2 and 7 days of exposure and 0 and 100 ng/L for 1 day of exposure (dimethyl sulfoxide with a final concentration of 1:2,500 v/v water). Each E2 treatment had two replicates with 90 larvae for each chamber. At day 1, 2 and 7 of the exposures, triplicate samples (30 larvae/sample) from each chamber respectively were collected, flash-frozen in liquid nitrogen, and stored in liquid nitrogen until RNA extraction.

All experiments were conducted with a 16 h light-8 h dark photoperiod and at 24 ± 1°C, except for reproduction tests, when the water temperature was raised to 27 ± 1°C to stimulate spawning. The fish were fed *Artemia nauplii* (<24 h after hatching) twice a day.

2.4 *Biological assessment*

1) Reproduction tests

At the end of the chronic exposures, sexual counterparts (2 months posthatch) for the E2-exposed fish were selected from the brood stock. Each mating pair of the E2-exposed fish and a sexual counterpart was put into a reproduction test chamber (1L) with circulated carbon-filtered tap water without E2. Fecundity (spawned eggs/pair/d) and fertility (fertilized eggs/total spawned eggs) were examined daily for each mating pair during 7 consecutive days and calculated for each E2 treatment group. At the end of the reproduction tests, the E2-exposed fish were separated from their sexual counterparts for gonad histology and genotypic sex determination.

2) Gonad histology

The E2-exposed fish were sacrificed, and their body lengths and weights were measured. Gonads were then sampled and weighed for the gonadosomatic index (GSI) calculation. The gonads were immersed in Bouin's fixative (Muto Pure Chemicals Co., Ltd., Tokyo, Japan) overnight. The gonads were then dehydrated and embedded in paraffin wax using a Shandon Excelsior ES (Thermo Fisher Scientific Co., Waltham, MA, USA). Serial longitudinal sections (5 μm of thickness) were prepared using a microtome. The sections were stained with hematoxylin and eosin, mounted with Mount-Quick (Daido Sangyo Co., Ltd., Kobe, Japan), and examined under a light microscope.

3) Genotypic sex determination

To identify sex reversal by the chronic exposures of E2, genotypic sex, XX or XY, was determined by detecting the presence of two major sex determination/differentiation genes (DMY/DMRT1bY and DMRT1) [17-19]. For this, we collected a piece of the caudal fin of all individuals for each treatment and extracted their total DNA using the DNeasy ® Tissue and Blood Kit (Qiagen Inc., CA, USA). A PCR analysis was then performed to detect both DMY and DMRT1 using a primer set, PG17.5 and PG17.6 (the nucleotide sequences: CGGGTGCCCAAGTGCTCCCGCTG and GATCGTCCCT CCACAGAGAAGAGA, respectively) according to Matsuda et al. [19]. The PCR products were analyzed electrophoretically with DNA 7500 Nano LabChip Kit using Agilent Bioanalyzer 2100 (Agilent Technologies, CA, USA).

2.5 *Microarray experiment*

1) Total RNA extraction and cRNA preparation

Each frozen sample was ground with a mortar and pestle in liquid nitrogen. Total RNA was isolated from the homogeneous powder using the Qiagen RNeasy Lipid Tissue Midi Kit (Qiagen) following procedures recommended by the manufacture. The quantity and purity of the total RNA were examined photometrically by 260nm/280nm and 260nm/230nm ratios using a NanoDrop ND1000 spectrophotometer (Nano Drop Technologies, DE, USA) and electrophoretically with RNA 6000 Nano LabChip Kit using Agilent Bioanalyzer 2100 (Agilent Technologies). Only RNA samples with RNA Integrity Number (RIN) values above 9.0 were used for further cRNA preparations.

cRNA was prepared from the total RNA using the Quick Amp Labeling Kit (Agilent Technologies) following procedures recommended by the manufacturer. Briefly, 500 ng of the total RNA was reverse transcribed to cDNA followed by synthesis of cRNA incorporated with cyanine 3 (Cy3)-labeled nucleotide. cRNA was then purified using RNeasy mini columns (Qiagen). The quality of the cRNA samples was verified by total yield of the cRNA and the incorporation rates of Cy3 calculated based on the spectrophotometric measurement using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies).

2) Microarray analysis

A medaka DNA microarray containing 36,398 genes was designed by the Chemical Evaluation and Research Institute (CERI) (Tokyo, Japan). The custom medaka DNA microarray with a 4 x 44K format was developed by Agilent Technologies based on the design. cRNA was fragmented to an average size of 100 bp. Hybridization was performed with single cRNA derived from one biosource (one-color hybridization). Hybridization, washing, and scanning were carried out following standard procedures (Agilent Technologies). The data used in further analyses consisted of local background-corrected median intensities that were greater than 2.6 standard deviations above the local mean background. The median of the selected raw data was calculated for each array. Normalization to the median was then applied to the raw intensities across the arrays. Differentially expressed genes were selected by a t-test with a p-value<0.05 for at least one treatment group at day 1, 2 or 7, and based upon a change greater than two-fold in comparison with the control group. Selected genes were categorized based on a biological process using the Gene Ontology (GO) (http://www.geneontology.org/) and UniProtKB (http://www.uniprot.org/uniprot/). Dose-response and time-dependent relationships in gene expression levels of the categorized genes were analyzed using the cumulative chi-squared method [20]. All microarray data have been submitted to the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) platform (http://www.ncbi.nlm.nih.gov/geo/) under platform number: GSE17633.

3. Results

3.1 *Chronic effects of E2 on medaka*

We first evaluated chronic effects on medaka exposed to E2 at 1, 3, 10, 30, and 100 ng/L for 60 days posthatch. Table 1 summarizes the results of the chronic experiments. From the observation of external sexual characteristics at 60 d posthatch, all fish in the 100 ng/L treatment group were phenotypically female (Table 1). PCR analyses of the caudal fins were then conducted to determine genotypic sex of all individuals in each E2 treatment using PG17.5 and PG17.6 primers (see "Materials and Methods"). With these primers, only one PCR product (DMRT1 fragment) was detected for the XX (female) genotype, whereas two PCR products (DMY and DMRT1 fragments) were generated for the XY (male) genotype. Of the phenotypic females, two (out of 13 individuals) and eight (out of 20 individuals) were determined as genotypic males in 30 ng/L and 100 ng/L treatment groups, respectively (Table 1). Sex reversal rates, (no. of sex-reversed males)/(no. of total genotypic males), were calculated to be 22.2% for the 30 ng/L treatment group and 100% for the 100 ng/L treatment group. No sex reversal was detected at or less than 10 ng/L of E2 treatment.

Gonad histology showed that there was no adverse effect in E2-treated males at or less than 3 ng/L or in all female individuals, whereas intersex gonads (testis-ova: oocytes in the testes) (Fig. 1) were present in one male of the 10 ng/L group and three males of the 30 ng/L group, respectively (Table 1).

The mean fecundities and fertilities of the E2-exposed males paired with unexposed females decreased significantly (ANOVA followed by a pair-wise t-test; Table 1) at 10 and 30 ng/L of E2. The mean fertilities of the E2-exposed females paired with unexposed males were significantly reduced in all E2-treatment groups except in the 30 ng/L group, whereas the mean fecundities of the E2-exposed females decreased significantly only in the 100 ng/L group (Table 1). No statistically significant difference was observed in GSIs of the E2-exposed females for all E2 treatment groups, whereas those of

Table 1. Chronic effects on medaka exposed to various concentrations of E2 for 60 days posthatch.

E2 concn. (ng/L)[1]	n	Phenotype sex			Gonadal histology (No. of fish)			GSI (%)[2]		Fertility (%)[2]		Number of eggs [2] (Number of eggs/pair/day)	
		male	female	female									
		Genotype sex			Testis	Ovary	Testis-ova	Male	Female	Male	Female	Male	Female
		XY	XX	XY									
0	21	10	10	0	10	10	0	0.019±0.0066	0.12±0.012	96.8±1.7	96.6±1.2	37.5±8.6	27.7±8.6
1	22	10	12	0	10	12	0	0.011±0.0035**[3]	0.11±0.029	82.7±11.6	91.9±4.9*	33.7±11.6	35.8±10.5
3	21	8	13	0	8	13	0	0.015±0.0045	0.12±0.031	77.5±26.7	89.8±5.1**	35.1±9.7	31.7±9.4
10	25	13	12	0	13	12	1	0.012±0.0036*	0.11±0.02	63.4±34.5*	78.4±19.4*	26.2±9.6*	25.7±4.5
30	20	7	11	2	7	13	3	0.013±0.0031*	0.12±0.043	55.7±32.1*	93.3±8.1	25.1±9.8*	20.3±10.4
100	20	0	12	8	0	20	ND[4]	ND	0.13±0.056	ND	69.8±26.4**	ND	17.5±13.2*

[1] Nominal concentrations.

[2] Data expressed as mean ± standard deviation.

[3] Significantly different compared with controls (*p<0.05, **p<0.01)

[4] ND: not detemined because all genotypic males were sex-reversed to phenotypic females.

the E2-exposed males decreased significantly in all E2 treatment groups except in the 3 ng/L group, although there appeared to be no dose-dependent effect (Table 1). Based upon the chronic experiments, 3, 30 and 100 ng/L of E2 were used in the exposure experiments for the DNA microarray analysis as low, medium, and high doses, respectively.

3.2 *Reproducibility of gene expression profiles in larval medaka*

For DNA microarray analysis, we employed larval medaka instead of mature fish, because critical steps in sex determination occur during the larval stage [21]. Because adult fish have normally been used for DNA microarray analyses [22-26], we first evaluated the validity of using larval medaka for DNA microarray experiments. The mean yields of total RNA from each sample (containing 30 larvae) were 46.5 μg, 42.6 μg and 58.1 μg for 1, 2 and 7 days of exposure, respectively. The 260 nm/280 nm ratios and RIN values of the total RNA samples used for microarray analyses ranged from 1.98 to 2.13 and from 9.7 to 10.0, respectively, indicating a high quality of the total RNA samples from larval medaka. We next examined reproducibility of gene expression in larval medaka. Table 2 shows correlation coefficients of gene expression patterns among controls for 1, 2 and 7 days posthatch. The correlation coefficient values ranged from 0.88 to 1.00 for control samples (an average value: 0.95), indicating that variances between mRNA expression profiles of larval medaka were small. These results demonstrate that highly reproducible gene expression data can be obtained from larval medaka under the experimental conditions used in this study.

3.3 *Gene expression profiles in response to E2 exposures*

1) Overview of functional categories of genes regulated by E2

Transcriptional changes were examined for the E2-exposed larvae using the medaka microarray. Differentially expressed genes were selected for each treatment based on the t-test and magnitude of change compared with the control group. Based

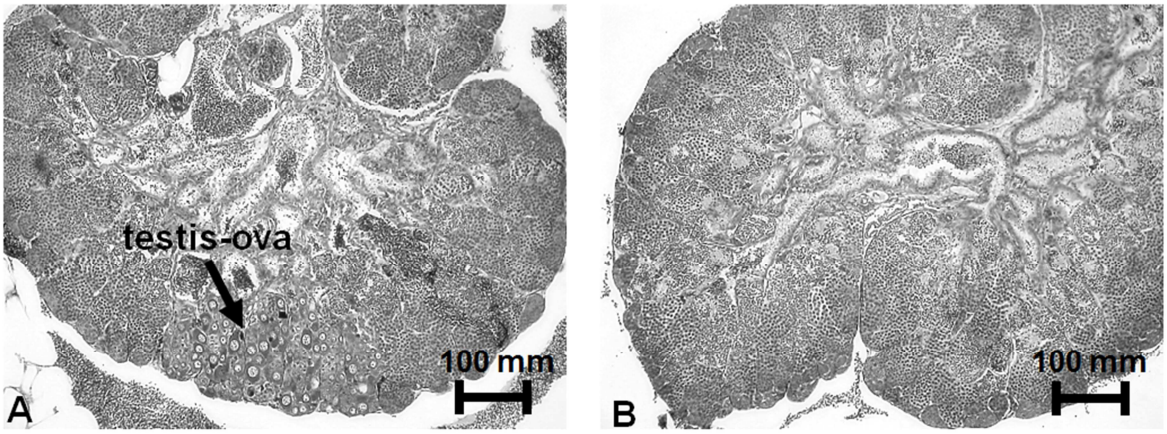


Figure 1. Effects of 17β-estradiol (E2) treatment on the testis in medaka. (A) Testis-ova observed in male medaka exposed to 100 ng/L E2 for 2 months posthatch. (B) Normal testis in control male medaka.

Table 2. List of correlation coefficients among control samples.

		1day[1]			2day[2]			7day[3]		
	No.	1	2	3	1	2	3	1	2	3
1day	1	1	0,97	0,98	0,96	0,97	0,98	0,92	0,87	0,88
	2		1	1,00	0,98	0,99	0,99	0,93	0,92	0,92
	3			1	0,98	0,99	0,99	0,93	0,92	0,93
2day	1				1	0,98	0,97	0,92	0,94	0,95
	2					1	0,99	0,93	0,92	0,93
	3						1	0,93	0,91	0,91
7day	1							1	0,91	0,92
	2								1	1,00
	3									1

[1] Control samples at day 1
[2] Control samples at day 2
[3] Control samples at day 7

on these criteria, 339 and 105 genes were determined as up- and down-regulated genes by the E2 treatment, respectively (up- or down-regulated in at least one exposure group) (Supplementary Table 1). Fig. 2 shows the numbers of up- and down-regulated transcripts for 1, 2 and 7 days exposure at different E2 concentrations. The numbers of up-regulated transcripts in the E2-exposed larvae dramatically increased at day 7, suggesting that the E2 treatments induced various physiological changes at and after day 7.

The up- and down-regulated genes were then categorized into functional groups (Fig. 3). Approximately 60% of both the up- and down-regulated genes were unknown for functions. Approximately 1/4 of the up- or down-regulated genes with known functions were in the "metabolism" category. The other 20% were assigned to "regulation of biological process" or "stress response", respectively.

2) Genes related to sexual differentiation and development

Table 3 summarizes expression profiles of representative genes related to sexual differentiation and development.

In the "hypothalamic-pituitary-gonadal (HPG) axis" category a few genes, such as brain aromatase, were significantly induced and exhibited a dose-dependent increase with E2 treatment at day 7. Expressions of estrogen receptor (ER)-related genes and androgen receptor (AR) α showed a dose-response decrease with E2 treatment. The remaining genes in the HPG category did not exhibit a dose-response relationship with E2 exposure. Time-dependent changes (1 to 7 days) in gene expression were observed for approximately 50% of the genes in this category such as 17-β hydroxysteroid dehydrogenase type 3 and nuclear receptor coactivator 7. Exposure to 100ng/L of E2 appeared to attenuate the time-dependent changes in several genes, particularly 17-β hydroxysteroid dehydrogenase type 3, activin A receptor type II, and nuclear receptor coactivator 7.

In the "oogenesis" category, vitellogenins (VTGs), choriogenins (CHGs) and L-SF precursors (L-SFs) were strongly induced with 30 ng/L and 100 ng/L of E2. The expression levels of these genes increased sharply with the E2 treatments in a dose-response and time-dependent manner. Time-dependent increases in gene expressions were observed for a factor in the germ line α (FIGα) and zona pellucida (ZP)-domain proteins, although these genes did not exhibit dose-dependent changes with E2 treatment.

Protamine in "spermatogenesis" did not show either a dose-response or time-dependent change when treated with E2.

3) Other genes with known functions

Table 4 summarizes expression profiles of the up- or down-regulated genes with known functions besides the genes related to sexual differentiation and development.

In the "regulation of biological process" category we found the up-regulated genes mainly belonged to three groups; "p53 signaling", "apoptosis", and "development and growth". Most of these genes exhibited a dose-dependent increase in gene expressions, although the increase was not as dramatic as VTGs or CHGs (Table 3). Approximately 70% of the genes in the "regulation of biological process" and "metabolism" categories showed a time-dependent decrease in gene expression for controls, suggesting that these genes may play important roles during the early larval stage. The E2 treatments appeared to compensate for the decrease in expression of these genes. Consequently, these genes had similar expression levels until day 7.

Many genes encoding heat shock proteins (HSPs) were down-regulated by the E2 treatments, showing a dose-response decrease in expression levels. Most of these genes exhibited a time-dependent increase in gene expression for controls. The E2 treatments, in this case, appeared to attenuate the time-dependent induction of the HSPs.

4. Discussion

It is well known that exogenous E2 adversely affects sexual development and reproduction of wildlife including fish [1]. Various mechanistic studies have been conducted on the

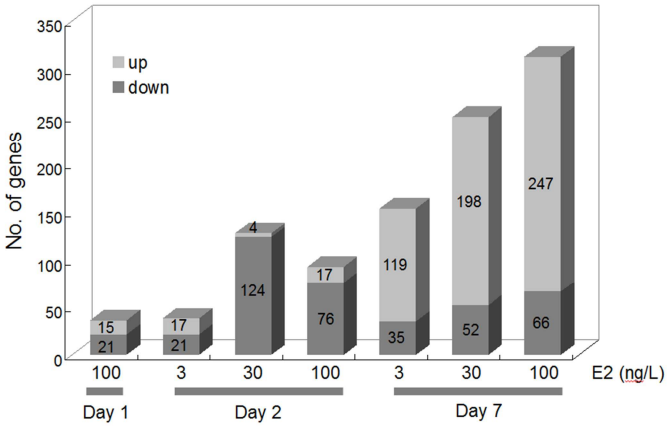


Figure 2. Numbers of up- and down-regulated genes in larval medaka exposed to 17β-estradiol.

endocrine disrupting effects of E2 mainly using fish species [2, 23, 27, 28]. Most of the studies, however, focused on elucidating the effects on the HPG axis by measuring the expression levels of several biomarker genes [11, 27-30]. A more integrated approach such as toxicogenomics is needed for a better understanding of endocrine disrupting phenomena. In this paper, we conducted a comprehensive analysis of the physiological effects of E2 on medaka using the medaka DNA microarray.

Larval medaka, instead of adult fish, were used for both chronic experiments and the transcriptomic analysis in this study, although adult fish have normally been used in various studies on endocrine disruptors [2, 4, 5, 24-26, 31]. Reasons why adult fish have been used are as follows: 1) the morphological sex is easily distinguished by observing secondary sexual characters [32], so that exposure experiments can be done separately for females and males; 2) internal organs can easily be separated compared with larvae, so that assays for biomarkers such as hepatic VTGs [11] can easily be conducted. One of the disadvantages of using adult fish is difficulty in evaluating effects of chemicals on development and sexual differentiation. Because organisms at early developmental stages are susceptible to endocrine disruptors [33], it is crucial to detect transcriptomic changes in embryonic, larval or juvenile stages to understand the mechanisms of endocrine disruptions. In fact, several key genes are expressed in late embryonic and early larval stages: a factor in the germ line α (Figα) at 1 day after hatching; aromatase at 4-10 days after hatching; DMY at embryonic stage 36 [34-36]. Thus, we decided to expose larval medaka to E2 immediately after hatching for chronic and microarray experiments.

We first conducted chronic experiments to determine E2 concentrations for microarray experiments. The full life cycle test by Seki et al. [33] showed that E2 caused reproductive impairment and feminization to medaka even at as low as 10 ng/L. Similar results were obtained in our study, although our experimental design was slightly different. In the work by Seki et al. [33], mating pairs were selected from the same E2-treatment group and exposed to E2 at the same concentrations for 30 days. The reproductive impairments might be caused during sex differentiation, gonad development, reproductive behavior or fertilization. With this experimental design, it might be difficult to clarify which sex is susceptible to E2 and at which developmental stages reproductive impairments were caused. Thus, in our study, the E2-exposed fish were paired with a non-exposed counterpart. Reproduction tests were then conducted without E2 exposure. Significantly lower fertility and egg production in the E2-exposed male (Table 1) suggest that E2 exposure might affect testicular development and reproductive behavior of male medaka. In female medaka, oogenesis, not oocyte development, could be adversely affected by E2 exposure, suggested by significantly lower fertility but no effect on GSIs (Table 1). Results also implied that a major part of adverse effects on both males and females were caused during development, not during reproduction tests.

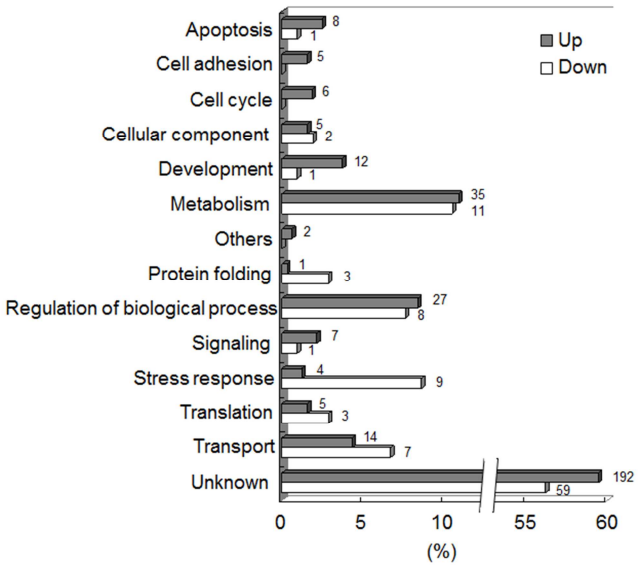


Figure 3. Functional categories of the up- and down-regulated genes in larval medaka exposed to 17β-estradiol based on the Gene Ontology (http://www.geneontology.org/).

Another different experimental design used in this study was genotype sex determination using PCR. With the primers used, only one PCR product (DMRT1 fragment) was detected for the XX (female) genotype, whereas two PCR products (DMY and DMRT1 fragments) were generated for the XY (male) genotype [19]. The sex of medaka is normally determined by observing secondary sexual characters (anal fins). Sex ratios have been compared by statistical analyses in many endocrine disruption studies [2-4, 18, 31]. Sex reversal, however, can only be confirmed by genotypic sex determination. Our results indicate that sex reversal occurred at ≥ 30ng/L of E2, higher than the E2 concentrations causing reproductive impairment (Table 1).

Changes in gene expression in larval medaka exposed to E2 were investigated using the medaka DNA microarray containing 36,398 features. Difficulty in sampling individual tissues from larvae led us to choose whole-body sample preparations. We first examined the variance in mRNA expression levels between pairs of control samples (30 larvae/sample) at day 1, 2 and 7. The average correlation coefficient among control samples was 0.95 (Table 2), indicating that mRNA expression profiles obtained by microarray analysis of medaka larvae are highly reproducible. The reproducibility makes larval medaka an ideal test animal for evaluating effects of chemicals, because of the small size of larvae compared with adult fish: for exposure, only a small space is required, and variances among samples should be smaller than for adults because one sample contains 30 larvae. Differentially expressed genes were then determined in the E2-exposed larvae for 1, 2, and 7 days. For 1 and 2 days of exposure, only 15 and 17 genes, respectively, were determined as induced genes even in the 100 ng/L E2 treatment groups (Fig. 2), suggesting that larvae must be exposed for at least 7 days for comprehensive analysis of the responses of genes to various chemicals.

Table 3. Expression profiles of genes related to sexual differentiation and development in larval medaka exposed to 17b-estradiol (E2).

Probe ID	Dose-response at day 7[1]				Time-dependence[2]						Description
	E2 (ng/L)			Dose-dependent increase (I) or decrease (D)[3]	Control		Time-dependent increase (I) or decrease (D)[3]	E2 100ng/L		Time-dependent increase (I) or decrease (D)[3]	
	3	30	100		2 day	7 day		2 day	7day		
"Hypothalamic-pituitary-gonadal axis"											
NP418483	1.27*	1,14	1.30*	I	1,26	0.76*	D	0,96	0,88	-	Activin A receptor type II
NP835138	1,03	1,08	1,45	-	1,16	0,82	-	0,91	0,97	-	Activin/inhibin b A chain protein
TC43331	0,90	0,63	0,78	-	1,00	1,54	-	0,86	1,05	-	Gonadotropin-releasing hormone receptor 1
TC43341	1,19	1,20	1,38	-	1,02	0.62*	D	0,96	0,75	-	Gonadotropin-releasing hormone receptor 2
TC43335	0,58	0,65	0,80	-	0,94	1,02	-	0,87	0,72	-	Gonadotropin-releasing hormone receptor 3
AU171640	1,22	1,14	1,38	-	1.16*	0,80	-	0,95	0,84	-	Neuropeptide Y
TC44068	1,06	0,93	0,98	-	0,96	1,03	-	0.85*	0,85	-	11b-hydroxysteroid dehydrogenase type 3
BJ733965	1,24	1,36	1,29	-	0,95	0.66*	D	0,94	0,98	-	17b-hydroxysteroid dehydrogenase type 3
TC44446	1,25	1,31	1,42	-	1,13	1,03	-	1,08	1,15	-	20b-hydroxysteroid dehydrogenase
AM149489	1,22	1,33	1,48	-	1,07	1,03	-	1,15	1,29	-	homolog to 20b-hydroxysteroid dehydrogenase
TC59351	1,02	1,06	1,25	-	0,88	0,95	-	1,14	0,91	-	3b-hydroxysteroid dehydrogenase
NP828099	1.39*	2.17*	4.41*	I	1,94	0,76	-	1,10	0,98	-	Brain aromatase (Cytochrome P450 19A2)
TC43315	1,15	1,23	1,07	-	0,96	1.61*	I	1,00	1.49*	I	Cytochrome P450 19A1
NP855454	1,18	1,06	1,24	-	1,04	0.54*	D	1,08	0,57	D	Cytochrome P450 11b
NP418747	1,02	1,35	2,57	I	1,19	1,14	-	1,02	2,40	-	Estrogen receptor
NP423661	2,24	1,41	1,56	-	1,23	1,52	-	0,77	1,17	-	Estrogen receptor b
AU169990	0,54	0.26*	0,92	-	0,91	1,03	-	0,99	1,23	-	Estrogen-related receptor a
TC43574	0,56	0.32*	0.23*	D	0,94	3.85*	I	1,02	0,82	-	Nuclear receptor coactivator 7 (140 kDa estrogen receptor-associated protein)
TC56094	0,50	0.21*	0.11*	D	0,81	4.07*	I	1,08	0,67	-	Nuclear receptor coactivator 7 (140 kDa estrogen receptor-associated protein)
NP863975	0.47*	0.57*	0.27*	D	0.62*	5.10*	I	0,89	2,18	-	Androgen receptor a
TC56245	0,99	1,03	0,95	-	0,94	0,96	-	0,83	1,04	-	Androgen receptor b
"Oogenesis"											
TC53300	4,77	144.51*	439.96*	I	66.21*	13,75	I	4.72*	91.52*	I	Choriogenin H
TC57172	2,03	152.66*	1383.36*	I	4,19	4,03	-	7.86*	449.77*	I	Choriogenin Hminor
TC55876	1,83	81.07*	741.72*	I	2,93	2,13	-	6.81*	394.28*	I	Choriogenin Hminor
TC54002	1,56	116.85*	1075.94*	I	5,48	3,35	-	7.70*	475.67*	I	Choriogenin Hminor
BJ915456	0,49	12,10	94.38*	I	3,01	4,24	-	2,64	204.90*	I	Choriogenin Hminor
TC59626	2,01	2.20*	2.24*	I	1,33	0.71*	D	1,19	1,44	-	homolog to Choriogenin Hminor

TC52879	1,04	1,45	2.14*	I	1,02	1,10	-	1,17	2.04*	I	homolog to Choriogenin Hminor
TC61048	1,25	5,66	107.15*	I	0,91	0.52*	D	1,26	81.63*	I	homolog to Choriogenin Hminor
TC54239	1,07	120.94*	708.48*	I	3,34	3,48	-	11.17*	477.58*	I	homolog to Choriogenin L
BJ923281	3,32	50,55	3670.37*	I	0,24	0,71	-	1,65	5502.26*	I	homolog to Choriogenin L
TC52912	1,59	20,68	70.60*	I	1,32	1,33	-	2.99*	52.21*	I	similar to Choriogenin H
TC57609	1,48	2.03*	2.35*	I	0,97	0,92	-	1,07	2.27*	I	Vitellogenin
TC52966	0,87	194,70	3007.41*	I	1,00	1,77	-	3,36	4582.71*	I	homolog to Vitellogenin 1
TC53352	1,04	1,54	24.58*	I	0,95	0.36*	-	0,90	10,90	-	homolog to Vitellogenin 1
TC58422	1,23	85,45	2427.07*	I	1,12	1,25	-	1,58	2409.29*	I	homolog to Vitellogenin 1
TC59208	0.58*	79,01	2734.28*	I	2,01	1,62	-	1,71	2279.88*	I	homolog to Vitellogenin 1
TC56202	2.09*	266,37	5208.38*	I	2.68*	2.41*	I	2,16	3619.01*	I	homolog to Vitellogenin 1
TC59433	2,12	98.65*	724.88*	I	3,85	3,27	-	8.42*	426.26*	I	homolog to Vitellogenin 1
TC59419	2,79	143.59*	606.12*	I	8,35	2.68*	-	5.01*	90.75*	I	homolog to Vitellogenin II
TC60262	2.85*	2.95*	4.26*	I	0,92	0.30*	D	0,92	1,25	-	L-SF precursor
TC53637	3,91	115.57*	306.04*	I	29,97	20,49	-	5.09*	90.89*	I	L-SF precursor
TC56827	1,94	184.37*	1667.81*	I	1,72	1,42	-	10.24*	660.27*	I	L-SF precursor
BJ918864	0,72	130.91*	1076.61*	I	1,64	1,76	-	10.63*	629.76*	I	L-SF precursor
TC59032	0,99	227.01*	1479.28*	I	1,48	2,46	-	13.65*	743.48*	I	homolog to L-SF precursor
BJ908450	1,21	192.71*	1507.28*	I	2,65	2,72	-	10.87*	630.97*	I	homolog to L-SF precursor
TC60472	1,04	109.27*	877.79*	I	2,34	2,81	-	9.66*	545.81*	I	homolog to L-SF precursor
BJ911644	2.04*	2.37*	2.45*	I	1,21	0,91	-	2,63	2.94*	I	homolog to L-SF precursor
TC45512	1,39	2,06	1,07	-	1,39	10.34*	I	1,11	9.57*	I	FIGa
TC43292	1,69	2,18	0,97	-	0,64	119.23*	I	1,90	212.80*	I	ZPAX
TC43334	1,84	2,59	0,93	-	0,99	28.57*	I	0,86	24.37*	I	ZPB domain containing protein
TC43355	1,53	2,24	0,94	-	2,72	283.52*	I	1,75	217.86*	I	ZPC domain containing protein 2
TC43477	1,68	2,65	1,10	-	1,27	46.15*	I	2,20	79.43*	I	ZPC domain containing protein 2
AM145452	2,55	3,57	1,74	-	1,24	2,97	-	0,69	2,19	I	ZPC domain containing protein 3
TC43321	1,46	2,03	0,99	-	2.80*	54.02*	I	4,93	48.22*	I	ZPC domain containing protein 4
BJ893265	2,97	3,70	3,15	-	1,06	1,06	-	1,27	4.15*	I	ZPC domain containing protein 5
TC43318	1,90	2,62	0,97	-	0,53	29.96*	I	1,76	65.95*	I	ZPC5
"Spermatogenesis"											
NP418658	0,80	0,92	1,51	-	1,53	1,40	-	0,14	0,34	-	Protamine

[1]Gene expressions at day 7 expressed as the fold change compared with controls at day 7.

[2]Time-dependent changes in gene expressions expressed as the fold change compared with control or the E2 100ng/L group at day 1.

[3]Dose- or time-dependent changes in gene expressions were evaluated using the cumulative chi-squared analysis. "I" and "D" indicate significant dose- or time-dependent increase and decrease ($p<0.05$), respectively. "-" indicates no significant dose- or time-dependent change.

*Significantly different ($p<0.05$) compared with controls at day 7 (dose-response) and with control or the E2 100ng/L group at day 1 (time-dependence), respectively.

Table 4. Expression profiles of genes with known functions in larval medaka exposed to 17b-estradiol (E2).

Probe ID	Dose-response at day 7[1]				Time-dependence[2]						Description
	E2 (ng/L)			Dose-dependent increase (I) or decrease (D)[3]	Control		Time-dependent increase (I) or decrease (D)[3]	E2 100ng/L		Time-dependent increase (I) or decrease (D)[3]	
	3	30	100		2 day	7 day		2 day	7day		
"Regulation of biological process"											
(a) p53 signaling											
TC43995	3.29*	4.55*	4.02*	I	0,77	0.24*	D	0,79	0,98	-	weakly similar to ankyrin repeat domain-containing protein 9
TC44618	3.86*	4.58*	4.14*	I	0,75	0.22*	D	0,88	0,96	-	weakly similar to ankyrin repeat domain-containing protein 9
TC45259	1.76*	1.76*	2.19*	I	1,38	0,67	D	1,15	1,13	-	BHLH protein DEC1b
TC46609	1,82	1,64	2.39*	I	1,32	0.48*	D	1.14*	1,00	-	similar to CBP/p300-interacting transactivator
TC46691	2.58*	2.08*	2.24*	I	1.72*	1,13	-	1,18	1,27	-	BHLH protein DEC1a
TC48131	1,65	2.03*	2,08	-	1,05	0.56*	D	1,09	1,18	-	similar to CCAAT/enhancer binding protein delta2
TC60219	1,18	1,37	4,48	I	1,06	0,78	-	0,89	4,15	I	weakly similar to apoptosis-stimulating of p53 protein 2
TC61529	1,72	2,16	1,40	-	0,98	1,12	-	1,15	1,47	-	similar to hypoxia-inducible factor 1 a
(b) apoptosis											
AM149239	2.18*	2.08*	1.94*	I	1,12	0,82	-	1,08	1.50*	I	similar to peroxisome proliferator-activated receptor g
AU178411	1,55	1.71*	2.09*	I	1,16	0,76	D	1,05	1,05	-	weakly similar to intestinal Muc 2-like protein
TC44023	1.79*	1.77*	2.18*	I	1,01	0.48*	D	0,98	0,99	-	homolog to ubiquitin
TC47349	1.64*	1.69*	2.03*	I	1,06	0.54*	D	0,97	0,99	-	homolog to ubiquitin
TC51915	2.15*	2.16*	2.23*	I	1,16	0.48*	D	1,06	0,98	-	similar to interferon regulatory factor 2 binding protein 2 isoform B
TC52486	1.82*	1.80*	2.15*	I	0,96	0.44*	D	0,95	0,95	-	homolog to ubiquitin
TC55411	2,20	2.28*	2.56*	I	1,21	0.57*	D	1,19	1,36	-	similar to caspase-9
TC55763	2.06*	2.33*	2.60*	I	1,26	0.65*	D	1,09	1,37	-	homolog of ring finger protein 36 isoform b
TC58172	2.13*	2.41*	2.52*	I	1,22	0.60*	D	1,22	1,39	-	weakly similar to caspase-9 precursor
(c) growth and development											
AM138341	1,56	1.88*	2.17*	I	1,01	0.37*	D	1,02	0,85	-	similar to HMG-box transcription factor 1
AU170803	2,92	5,07	2.55*	I	0,75	0,70	-	1,93	1,53	-	similar to zinc finger RNA-binding protein
AU178201	2.06*	1.90*	1.89*	I	0,92	0.53*	D	0,89	0,96	-	similar to connective tissue growth factor precursor
BJ011762	1,59	1,63	2.05*	I	1,11	0.30*	D	0,97	0.51*	D	similar to keratin

BJ025587	1,32	1,97	2,70	-	0,99	0,97	-	1,06	2,61	-	TUDOR
BJ708117	7,36	3.60*	7.30*	I	1,89	0,90	-	0,69	2.50*	I	similar to growth-arrest-specific protein 1 precursor (GAS-1)
BJ736582	1.56*	1.54*	2.04*	I	1,15	0.4*	D	1,09	1,05	-	similar to kinetochore protein Spc24
BJ880895	2,02	1,93	1,95	I	1,04	0,79	-	0,80	1.57*	I	similar to growth arrest and DNA- damage-inducible protein GADD153
NP1445793	2.57*	1,97	1,69	-	1.64*	1.97*	I	1,09	1,52	-	BHLH protein DEC2
TC43357	1.94*	2.08*	2.18*	I	1,61	1,09	-	1,14	1.76*	I	HoxC10a
TC45582	1,96	2.17*	2.62*	I	1,18	0.53*	-	1,25	1,79	-	homolog to Growth arrest and DNA-damage-inducible protein GADD45 g
TC45587	2,68	3.14*	9.05*	I	1,77	0.30*	D	1,53	2,09	-	similar to insulin-like growth factor binding protein-1 (IGFBP1)
TC48028	1,21	1,37	2.35*	-	0,91	0.38*	D	0,68	1,18	-	similar to protein Jumonji
TC51745	2.24*	2.55*	2.74*	I	0,98	0.40*	D	1,14	1,11	-	similar to cyclin-G2
TC52709	1.71*	1.90*	2.03*	I	1,23	0.66*	D	1,00	1,18	-	similar to development and differentiation enhancing factor-like 1
TC53134	2.13*	1.97*	2.01*	I	0,88	0.45*	D	0,93	0,90	-	similar to connective tissue growth factor precursor
TC53171	2.58*	3.31*	2.56*	I	1,04	0.72*	D	1,20	1.77*	I	similar to krueppel-like factor 11
TC57898	1.70*	1.77*	2.02*	I	1,35	0,94	-	0,82	1.81*	I	similar to growth arrest and DNA- damage-inducible protein GADD153
"Metabolism"											
cytochrome P450											
TC58366	1,29	1.56*	2.47*	I	1,04	0.27*	D	0,99	0,63	D	cytochrome P450 monooxygenase CYP2K1
TC58597	1,36	1.68*	2,15		1,01	0.31*	D	0,98	0,63	-	similar to cytochrome P450 2K5
TC43435	1,04	1.22*	1.13*	I	1,45	2.39*	I	1,15	1.82*	I	Cytochrome P450 3A
"Stress response" and "Metabolic process"											
heat shock protein											
AU169726	0,87	0,31	0,23		1,53	1,50	-	0,89	0.59*	D	stress-70 protein
TC52094	0,78	0.69*	0.40*	D	1,21	2.76*	I	1,27	1,07	-	similar to 47 kDa heat shock protein
TC53141	0,62	0,59	0.48*		1,17	1.90*	I	1.25*	0.75*	D	homolog to novel protein similar to heat shock protein 90-a
TC60176	0,74	0.54*	0.37*	D	0,88	1.91*	I	0,89	0,78	-	homolog to heat shock 70 kDa protein 5
TC61286	0,62	0,53	0.41*		1,05	2.10*	I	0,97	0,75	-	homolog to heat shock protein 90 a
TC62024	0,61	0.46*	0.46*	D	0,69	1.52*	I	0,89	1,10	-	heat shock cognate 71 kDa protein
TC49227	0,57	0.41*	0.47*	D	0,91	2.43*	I	0,92	1,14	-	homolog to 10 kDa heat shock protein mitochondrial

[1]Gene expressions at day 7 expressed as the fold change compared with controls at day 7.

[2]Time-dependent changes in gene expressions expressed as the fold change compared with control or the E2 100ng/L group at day 1.

[3]Dose- or time-dependent changes in gene expressions were evaluated using the cumulative chi-squared analysis. "I" and "D" indicate significant dose- or time-dependent increase and decrease ($p<0.05$), respectively. "-" indicates no significant dose- or time-dependent change.

*Significantly different ($p<0.05$) compared with controls at day 7 (dose-response) and with control or the E2 100ng/L group at day 1 (time-dependence), respectively.

Dose-response and time-dependent microarray analyses were then carried out to evaluate the E2 effects on larval development. Table 3 shows the expression profiles of the genes involved in sexual differentiation and development. The endocrine control of reproduction in medaka involves the hypothalamic-pituitary-gonadal axis [37]. The gonadotropin-releasing hormone (GnRH) is synthesized in the preoptic/hypothalamic area of the brain and stimulates secretion of gonadotropins [38]. In teleost fish such as medaka, gonadotropins have two forms: follicle-stimulating hormone (FSH) and luteinizing hormone (LH) [38]. FSH and LH are primarily involved in E2 biosynthesis and oocyte maturation, respectively. E2 plays critical roles in oocyte growth and is synthesized via the steroidogenesis pathway [39]. The steroidogenesis pathway includes various enzymes, such as 17β-hydroxysteroid dehydrogenase (HSD), 3β-HSD, 20β-HSD and aromatase. E2 is finally converted from testosterone, catalyzed by cytochrome P450 aromatase [39]. Medaka has two distinct genes for aromatase: 19A1 in the ovarian follicular layers and 19A2 in the brain, which play important roles in reproduction and neurogenic activity, respectively [40]. In this study, only a few genes, particularly brain aromatase 19A2, exhibited a significant dose-dependent induction by the E2 treatments (Table 3). The remaining genes in the HPG axis did not show either dose- or time-dependent responses to E2 exposure (Table 3). These results suggest that the estrogen treatments may affect neurogenic development in the brain, but appear not to modulate E2 biosynthesis, in line with the work on adult medaka by Zhang et al. [29].

As shown in previous studies [11, 27, 28,41], genes related to oogenesis, such as VTGs, CHGs and L-SFs, were strongly induced even at larval stages exposed to E2 (Table 3). VTGs and CHGs are precursor proteins of egg yolk and the egg envelope, respectively, that are synthesized in the liver [27, 28, 41]. VTGs, CHGs and L-SFs have been shown to be good biomarkers for environmental estrogens in fish, due to their sensitivities and specificities to E2 [11, 27, 28, 41]. Some of these genes were induced more than one thousand times relative to controls at 100 ng/L of E2 (Table 3), confirming that the VTGs and CHGs are the most sensitive biomarkers for estrogens. The induction of these genes exhibited both a dose- and time-dependent relationship and appeared to correlate with gonadal abnormalities and sex reversal rates. More studies, however, have to be done to elucidate the relationship between feminization phenomena and induction of the oogenesis-related genes.

Other genes related to the egg envelope, called the zona pellucida (ZP)-domain are specifically expressed in the oocyte [42]. Significant increases in gene expression of the ZP genes were observed in controls at day 7 (Table 3), indicating that gene expression during early oogenesis can be evaluated at day 7 posthatch. Although egg envelope formation seems to involve both CHG and ZP gene expressions [42], the E2 treatments did not cause any induction of the ZP genes (Table 3). In mammals, a factor in the germ line α (FIGα) plays a key role in the oocyte-specific expression of ZP genes [43]. In this study, FIGα did not exhibit dose-response gene expression, but showed a time-dependent increase in gene expression (Table 3). These results imply that FIGα may be involved in the expression of ZP genes, although further studies are necessary to clarify the interaction between FIGα and ZP genes in medaka.

Protamine is expressed at the initial stage of spermiogenesis in medaka and involved in the maintenance of the condensing state in sperm nuclei [44]. The gene expression of protamine did not show either dose- or time-dependent change with E2 treatments (Table 3). E2 exposure did not cause significant changes in gene expression of FIGα, ZPs or protamine even in mature fish [22], suggesting that E2 may not be involved in gene expression of these genes.

Interestingly, some of the ER-related genes were down-regulated by the E2 treatments (Table 3). It is known that ERs play important roles in defeminization of the male brain and, therefore, sexual behavior. Exposure of E2 during development (gestation and shortly after birth) has been suggested to suppress ER expressions in mammal brains through feedback inhibitions of its ligand, E2 [45]. ER expression levels increased when male rats were deprived of E2 after birth, whereas E2 supplements caused subsequent declines in ER expressions [46, 47]. These studies suggest that the expression levels of neural ERs might be influenced by endogeneous E2 levels during development. It was also reported that ERβ in the brain of goldfish was down-regulated by E2 exposure [48]. Taken together with our results, it is implied that excess administration of E2 may cause the repression of ER-related genes to maintain E2 homeostasis in larval medaka. Some genes encoding HSPs were also down-regulated, and exhibited a dose-response relationship to E2 (Table 4). ERs exist as monomers in the cytoplasm without their ligands, and form a multi-protein complex with HSP90 and HSP70 in the presence of E2 [49, 50]. In our study, the expressions of both HSPs and ER-related genes increased in a time-dependent manner for controls, which appeared to be attenuated by the E2 treatments (Table 3 and 4). It is, however, speculative to state that ERs and HSPs are co-regulated. More work needs to be done to understand the mechanism of HSP suppression by E2 exposure.

Besides the sexual development related gene expressions, many genes encoding p53 signaling, apoptosis, and growth and development were significantly induced even at as low as 3 ng/L of E2 (Table 4). The expression levels of these genes declined significantly later than day 2 for controls. The E2 treatments maintained these expression levels even at day 7 (Table 4). Some genes encoding p53 signaling, apoptosis, and development and growth are known to be involved in both development and stress response [51-55]. Genes related to the p53 signaling play important roles in cell cycle arrest [53, 54]. The p53 signaling pathway is required in embryonic development and DNA repair [53, 54]. Apoptosis is the process of programmed cell death that may occur in multicellular processes, including mammalian embryonic development [51, 52]. Apoptosis is also triggered when cells are damaged be-

yond repair (such as DNA damage) [51, 52]. GADD45, GADD153, and GAS1 related to DNA damage and cell cycle arrest are also known as genes associated with the regulation of growth during organogenesis in the rat fetus [55]. On the other hand, some of the cytochrome P450 enzymes convert E2 to 4-OH-E2, which undergoes metabolic redox cycling between hydroquinone and quinone, formed via semiquinone intermediates [56, 57]. The semiquinone is a free radical that can react with molecular oxygen to form oxygen reactive species, resulting in cell and DNA damage. Our results imply that the contribution of cytochrome P450-mediated oxidative stress to p53 signaling and apoptosis might be rather small, because only a few cytochrome P450s were significantly induced at high concentrations of E2. The p53 signaling and apoptosis at 3 and 30 ng/L of E2 could be caused by the stimulation of cell growth and development. Estrogens have been used as promoting agents for growth of cattle and sheep and also used for the estrogen therapy in Turner's syndrome [58]. In our experiments, significant increases were observed in body lengths and weights of the E2 30 ng/L group compared with controls (data not shown), suggesting that the larval growth might be promoted at low concentrations of E2.

In this study, we investigated physiological effects of medaka exposed to E2 using the combination of chronic experiments and transcriptomic analyses. Larval medaka instead of adult fish were used in this study to evaluate the effects of E2 on sexual differentiation and development. Larvae were fairly sensitive to E2 in gene expression, and the gene expression profiles were highly reproducible. Using medaka DNA microarray, the transcriptomic approach was conducted to determine the physiological effects of E2. Results suggest that E2 may affect not only the steroid receptor-mediated pathway, but also other physiological activities, such as apoptosis, cell growth and development. Because the gene expression profiles of the E2-exposed larval medaka were similar to those of mammals, larval medaka can be an excellent vertebrate model for a mechanistic study of endocrine disruption.

5. Supplementary material

Supplementary material regarding this manuscript is available online in the web page of JIOMICS.

http://www.jiomics.com/index.php/jio/rt/suppFiles/29/0

Acknowledgments

We would like to thank Chiyomi Imai and Rie Kimura for their assistance in maintaining medaka brood stocks and their contribution to the E2 exposure experiments.

References

1. T. Colborn, D. Dumanoshi, J.P. Mayers, Our Stolen Future, Dutton, New York, 1996.
2. I.J. Kang, H. Yokota, Y. Oshima, Y. Tsuruda, T. Yamaguchi, M. Maeda, N. Imada, H. Tadokoro, T. Honjo, Chemosphere. 47 (2002) 71-80.
3. M. Seki, H. Yokota, H. Matsubara, Y. Tsuruda, M. Maeda, H. Tadokoro, K. Kobayashi, Environ. Toxicol. Chem. 21 (2002) 1692-1698.
4. I.J. Kang, H. Yokota, Y. Oshima, Y. Tsuruda, T. Oe, N. Imada, H. Tadokoro, T. Honjo, Environ. Toxicol. Chem. 21 (2002) 2394-2400.
5. I.J. Kang, H. Yokota, Y. Oshima, Y. Tsuruda, T. Hano, M. Maeda, N. Imada, H. Tadokoro, T. Honjo, Environ. Toxicol. Chem. 22 (2003) 2438-2445.
6. K. Van den Belt, R. Verheyen, H. Witters, Arch. Environ. Contam. Toxicol. 41 (2001) 458-467.
7. L.T. Van der Ven, P.W. Wester, J.G. Vos, Environ. Toxicol. Chem. 22 (2003) 908-913.
8. G.T. Ankley, K.M. Jensen, E.A. Makynen, M.D. Kahl, J.J. Korte, M.W. Hornung, T.R. Henry, J.S. Denny, R.L. Leino, V.S. Wilson, M.C. Cardon, P.C. Hartig, L.E. Gray, Environ. Toxicol. Chem. 22 (2003) 1350-1360.
9. E. Diamanti-Kandarakis, J.P. Bourguignon, L.C. Giudice, R. Hauser, G.S. Prins, A.M. Soto, R.T. Zoeller, A.C. Gore, Endocr. Rev. 33 (2009) 293-342.
10. G.T. Ankley, K.M. Jensen, E.J. Durhan, E.A. Makynen, B.C. Butterworth, M.D. Kahl, D.L. Villeneuve, A. Linnum, L.E. Gray, M. Cardon, V.S. Wilson, Toxicol. Sci. 86 (2005) 300-308.
11. S.A. Heppell, N.D. Denslow, L.C. Folmar, C.V. Sullivan, Environ. Health Perspect. 103, Suppl 7 (1995) 9-15.
12. J.L. Dai, K.L. Burnstein, 10 (1996) 1582-1594.
13. D.V. Henley, K.S. Korach, Endocrinology 147, Suppl 6 (2006) S25-32.
14. J.T. Sanderson, Toxicol. Sci. 94 (2006) 3-21.
15. S.M. Dickerson, A.C. Gore, Rev. Endocr. Metab. Disord. 8 (2007) 143-159.
16. J.R. Snape, S.J. Maund, D.B. Pickford, T.H. Hutchinson, Aquat. Toxicol. 67 (2004) 143-154.
17. M. Matsuda, Dev. Growth Differ. 45 (2003) 397-403.
18. M. Kondo, I. Nanda, M. Schmid, M. Schartl, Sex. Dev. 3 (2009) 88-98.
19. M. Matsuda, Y. Nagahama, A. Shinomiya, T. Sato, C. Matsuda, T. Kobayashi, C.E. Morrey, N. Shibata, S. Asakawa, N. Shimizu, H. Hori, S. Hamaguchi, M. Sakaizumi, Nature. 417 (2002) 559-553.
20. C. Hirotsu, Biometrika. 73 (1986) 165-173.
21. D. Saito, M. Tanaka, Sex. Dev. 3 (2009) 99-107.
22. K. Kishi, E. Kitagawa, N. Onikura, A. Nakamura, H. Iwahashi, Genomics. 88 (2006) 241-221.
23. H. Iwahashi, K. Kishi, E. Kitagawa, K. Suzuki, Y. Hayashi, Environ. Sci. Technol. 43 (2009) 3913-3918.
24. K. Nakayama, N. Sei, Y. Oshima, K. Tashiro, Y. Shimasaki, T. Honjo, Mar. Pollut. Bull. 57 (2008) 460-466.
25. A. León, P.S. Wu, L.C. Hall, M.L. Johnson, S.J. Teh, Environ. Sci. Technol. 42 (2008) 962-969.
26. S. Yum, S. Woo, Y. Kagami, H.S. Park, J.C. Ryu, Comp. Biochem. Physiol. C Toxicol. Pharmacol. 151 (2010) 51-56.
27. C. Lee, J.G. Na, K.C. Lee, K. Park , Aquat. Toxicol. 61 (2002) 233-241.
28. X. Chen, V.W. Li, R.M. Yu, S.H. Cheng, Ecotoxicol. Environ. Saf. 71 (2008) 200-208.
29. X. Zhang, M. Hecker, J.W. Park, A.R. Tompsett, J. Newsted, K. Nakayama, P.D. Jones, D. Au, R. Kong, R.S. Wu, Aquat. Toxicol. 88 (2008) 173-182.
30. X. Zhang, M. Hecker, J.W. Park, A.R. Tompsett, P.D. Jones, J. Newsted, D.W. Au, R. Kong, R.S. Wu, J.P. Giesy, Environ. Toxicol. Chem. 27 (2008) 2504-2511.
31. M. Seki, S. Fujishima, T. Nozaka, M. Maeda, K. Kobayashi, Environ. Toxicol. Chem. 25 (2006) 2742-2752.
32. M. Kondo, I. Nanda, M. Schmid, M. Schartl, Sex. Dev. 3 (2009) 88-98.
33. M. Seki, H. Yokota, M. Maeda, K. Kobayashi, Environ. Toxi-

col. Chem. 24 (2005) 1259-1266.
34. R.S. Hattori, R.J. Gould, T. Fujioka, T. Saito, J. Kurita, C.A. Strüssmann, M. Yokota, S. Watanabe, Sex. Dev. 1 (2007) 138-146.
35. A. Suzuki, M. Tanaka, N. Shibata, Y. Nagahama, J. Exp. Zool. A Comp. Exp. Biol. 301 (2004) 266-273.
36. A. Kanamori, Mol. Reprod. Dev. 55 (2000) 31-36.
37. G.T. Ankley, R.D. Johnson. ILAR J. 45 (2004) 469-483.
38. Y. Oka. J. Neuroendocrinol. 21 (2009) 334-338.
39. K. Okubo, A. Shinomiya, N. Hirai, T. Kitano, A. Watanabe, K. Murata, A. Shimada, T. Deguchi, K. Niwa, in M. Kinoshita, K. Murata, K. Naruse, M. Tanaka (Eds), Medaka: Biology, Management, and Experimental Protocols (2009) Wiley-Blackwell. Iowa, 2009, pp. 67-99.
40. D. Saito, M. Tanaka. Sex Dev. 3 (2009) 99-107.
41. K. Murata, H. Sugiyama, S. Yasumasu, I. Iuchi, I. Yasumasu, K. Yamagami, Proc. Natl. Acad. Sci. U.S.A. 94 (1997) 2050-2055.
42. A. Kanamori, K. Naruse, H. Mitani, A. Shima, H. Hori, Gene 305 (2003) 35-45.
43. S.M. Soyal, A. Amleh, J. Dean, Development 127 (2000) 4645-4654.
44. Y. Shimizu, K. Mita, M. Tamura, K. Onitake, M. Yamashita, Int. J. Dev. Biol. 44 (2000) 195-199.
45. A.E. Kudwa, N. Harada, S.I. Honda, E.F. Rissman, J. Neuroendocrinol. 19 (2007) 767-772.
46. J.L. Temple, H.N. Fugger, X. Li, S.J. Shetty, J. Gustafsson, E.F. Rissman, Endocrinology. 142 (2001) 510-513.
47. A.E. Kudwa, C. Bodo, J.A. Gustafsson, E.F. Rissman, Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 4608-4612.
48. V.L. Marlatt, C.J. Martyniuk, D. Zhang, H. Xiong, J. Watt, X. Xia, T. Moon, V.L. Trudeau, Mol. Cell. Endocrinol. 283 (2008) 38-48.
49. W.B. Pratt, D.O. Toft, Endocr. Rev. 18 (1997) 306-360.
50. M.E. Wierman, Adv. Physiol. Educ. 31 (2007) 26-33.
51. K. Hardy, Rev. Reprod. 4 (1999) 125-134.
52. L.K. Cole, L.S. Ross, Dev. Biol. 240 (2001) 123-142.
53. J. Choi, L.A. Donehower, Cell. Mol. Life Sci. 55 (1999) 38-47.
54. J.B. Wallingford, D.W. Seufert, V.C. Virta, P.D. Vize, Curr. Biol. 7 (1997) 747-757.
55. W.D. Rees, S.M. Hay, N.C. Fontanier-Razzaq, C. Antipatis, D.N. Harries, J. Nutr. 129 (1999) 1532-1536.
56. J.G. Liehr, Environ. Health Perspect. 105 Suppl 3 (1997) 565-569.
57. B.T. Zhu, H. Allan. Functional role of estrogen metabolism in target cells: review and perspectives. Carcinogenesis. 19 (1998) 1–27.
58. G.B. Cutler Jr, J.L. Ross, Acta. Paediatr. Jpn. 34 (1992) 195-202.