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

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v3i2.137

Determining the C-Terminal Amino Acid of a Peptide from MS/MS Data

Jens Allmer

Molecular Biology and Genetics, Izmir Institute of Technology, Urla, Izmir, Turkey

Received: 26 April 2013 **Accepted:** 16 June 2013 **Available Online**: 15 July 2013

ABSTRACT

Proteomics is currently chiefly based on mass spectrometry (MS) which is the tool of choice to investigate proteins. Two computational approaches to derive the tandem mass spectrum precursor's sequence are widely employed. Database search essentially retrieves the sequence by matching the spectrum to all entries in a database whereas de novo sequencing does not depend on a sequence database. Both approaches benefit from knowledge about the enzyme used to generate the peptides. Most algorithms default to trypsin for its abundant usage. Trypsin cuts after arginine and lysine and thus the c-terminal amino acid is not known precisely and usually either of the two. Furthermore, 90% of protein terminal peptides may not end with either arginine or lysine and may thus contain any of the other amino acids. Here an algorithm is presented which predicts the c-terminal amino acid to be arginine, lysine or any other.

Here an algorithm, named RKDecider, to sort the c-terminal amino acid into one of three groups (arginine, lysine, and other) is presented. Although around 90% accuracy was achieved during data mining spectra for rules that determine the c-terminal amino acid, the implementation's (RKDecider) accuracy is a little less and achieves about 80%. This is due to the fact that the decision trees were implemented as a rule-based system for speed considerations. The implementation is freely available at: http://bioinformatics.iyte.edu.tr/RKDecider.

Keywords: mass spectrometry; proteomics; trypsin cleavage; database search; *de novo* sequencing; fragmentation analysis, c-terminal amino acid.

Abbreviations

MS: Mass Spectrometry; **Da**: Dalton; **m/z**: Mass to charge ratio.

1. Introduction

Proteomics analyzes the sequence, localization, modifications, and other parameters of proteins. Currently mass spectrometry is the tool of choice in proteomics [1]. Since peptides are easier to bring into gas phase than large proteins, bottom-up proteomics is widely applied [2]. In short, proteins are digested by a protease and the resulting peptides are fed into a mass spectrometer where their mass to charge ratio and/or their fragmentation spectra are measured. Usually trypsin is used for the purpose of digesting proteins since it leads to peptides of desirable size for mass spectrometry and since it cleaves highly specific and efficient [3,4]. Despite the widespread use of trypsin its proteolytic process has not been analyzed in depth, but has recently been modeled from mass spectrometric data [5]. Since mass spectrometry can only measure mass to charge ratios, the measurements need to be translated back to peptide sequences and then integrated into proteins. To derive the sequence of a fragmentation spectrum (MS/MS spectrum), generally two different methods are employed. The first is database search, where the MS/MS spectrum is compared to all sequences in a database to determine the best fit [6]. The other method is *de novo* sequencing which needs no other information than the MS/MS spectrum to assign a protein sequence [7]. Many algorithms for database search and *de novo* sequencing require the knowledge of the enzyme used for digestion. In case trypsin was used, it is not clear whether the c-terminal amino acid is arginine or lysine. However, for many purposes it would be beneficial to know the c-terminal amino acid precisely (see below).

Peptide fragmentation in gas phase leads to the cleavage of peptides into smaller fragments, a process for which many

***Corresponding author:** Jens Allmer, tel.: 00902327507310, fax: 00902327507303. Email Address: jens@allmer.de; address: Assoc. Prof. Dr. Jens Allmer, Molecular Biology and Gentics, Izmir Institute of Technology, Gulbahce Campus, Urla, Izmir, Turkey.

methods are available [7]. The fragmentation mechanism, even for a time tested method like collision induced dissociation (CID) [8] is not completely understood and has not been completely integrated into algorithms that assign amino acid sequences to MS/MS spectra, although statistics have been derived from larger amounts of spectra [9]. These fragments are important for database search and *de novo* sequencing algorithms to derive the sequence of the peptide precursor that gave rise to the MS/MS spectra. Among the resulting peaks there may be some which could facilitate the distinction of which amino acid terminates a peptide. Previously, it has been shown that a peak, resulting from 17 Dalton or 42 Dalton eliminated from the precursor ion, may be diagnostic for arginine [10].

Olsen and colleagues showed that the largest amount of peptides derived from a tryptic digest have either arginine or lysine at the c-terminus and conclude that non tryptic peptide assignments, done by some algorithms, should not be trusted [4]. They acknowledge, however, that not all peptides that result from a tryptic digest have a c-terminal arginine or lysine [4]. Taking into account that there are 20 possible amino acids terminating a protein, 90% of them are not arginine or lysine. Some studies use non-specific cleavage to increase the number of identified spectra but this is controversial and likely only increases the number of false positive identifications [4]. Instead of globally turning a search engine to non-specific cleavage for all spectra, such costly and dangerous operations can be performed on the basis of the decision whether the c-terminus of the peptide, that gave rise to the mass spectrum, is not tryptic. Thus only spectra which are not tryptic trigger searches with non-specific enzymatic cleavage settings. Additionally, it is of help to know the c-terminus so that peptide candidates in database search can either be pre-filtered with the knowledge of the c-terminal amino acid or the results can be evaluated in respect to proper sequence selection based on the terminal amino acid.

In order to reap these benefits an approach and its implementation to decide whether the c-terminus of peptides, underlying CID spectra from LTQ instruments, is tryptic and which amino acid is at the c-terminus, is presented here. 12 potential diagnostic losses for arginine and lysine were defined and data mining on about 8500 LTQ spectra was performed. From these, rules were derived that in data mining practice can distinguish between arginine, lysine, or other amino acids with an accuracy of about 90%. The practical implementation reaches an accuracy of about 80% since other limitations like speed and unknown charge state had to be taken into account.

2. Methods

2.1. *Spectral Dataset*

44 synthetic peptides, mostly derived from cytochrome-c and bovine serum albumin, were designed and ordered for a different purpose than this study and will be published elsewhere. The dataset had to be prepared in this way to have a ground truth for the analysis [11]. These peptides were directly injected into a Thermo Finnigan LTQ mass spectrometer and their collision induced dissociation fragmentation spectra were recorded. This resulted in 8447 MS/MS spectra with an average of 192 spectra per peptide. 42 peptides were tryptic while two (QVYQGCGV and YKELGFQG) were not and ended in valine and glycine, respectively. The complete list of peptides and the number of measured MS/MS spectra are available as Supplementary File 1. The resulting spectra were predominantly from singly charged precursors (5353); and spectra from doubly charged precursors (2616) as well as spectra from triply charged precursors (478) were less. 63% of the spectra are derived from a precursor terminating with lysine, 34% from a precursor terminating with arginine, and 3% had a precursor that terminated with a different amino acid.

2.2. *Diagnostic Fragments and Data Mining*

In order to determine whether a spectrum derives from a peptide precursor ending in arginine, lysine, or another amino acid, 12 parameters were defined. These parameters are relative losses of the precursor ion. A loss of 156.1 Da of the precursor ion, for instance, signifies the loss of arginine. In the following -156.1 shall signify a loss of 156.1 Da from the precursor. The chosen diagnostic fragments are listed and briefly explained in Table 1. Not all of the selected fragments have to exist in practice. Their absence can also be learned by the machine learning algorithms employed here and they can hence still be used as diagnostic fragments. The chosen diagnostic fragments are the following: -16, -17, -32, -33, -34, -42, -43, -57, -128.1, -129, -156.1, -175.

Orange Canvas [12] was used for all data mining and if not otherwise stated in the text, the default settings were used. For example in all cases 10 fold cross validation was used for learning and testing instead of the default which is only 5 fold.

2.3. *Software Implementation*

The implementation of the RKDecider software was written in Java™ using the Netbeans integrated development environment version 7.1. The implementation is available for download at http://bioinformatics.iyte.edu.tr/RKDecider.

3. Results

The recorded LTQ spectra were analyzed and their charge state was determined using the recorded precursor mass and the theoretically calculated mass. Since the sequence was known and there was no mixture this procedure guaranteed accurate charge determination. A window of +/- fragment tolerance (0.3 Da) around the peak with the diagnostic loss was extracted from all spectra and in case multiple peaks were found in the window their abundances were summed.

Table 1. In order to learn whether a spectrum derives from a peptide ending in arginine or lysine, a number of peaks, described as losses from the precursor ion, have been defined. Here the loss from the precursor ion is given in Daltons and the fragment is briefly explained.

Loss (Da)	Reasoning
-16	As -17 but accommodating for one hydrogen difference (accounting for possible dependence on the acidity of the solution)
-17	Loss of NH3 which is possible for arginine and lysine
-32	As -34 but accommodating for two hydrogen difference (accounting for possible dependence on the acidity of the solution)
-33	As -34 but accommodating for one hydrogen difference (accounting for possible dependence on the acidity of the solution)
-34	Loss of 2 NH3 which is only possible for arginine.
-42	As -43 but accommodating for one hydrogen difference (accounting for possible dependence on the acidity of the solution)
-43	Partial elimination of the terminal part of the side chain from arginine (C1N2H4).
-57	Full elimination of the terminal part of the side chain from arginine (C1N3H6).
-128.1	Molecular weight of lysine without water; elimination of the amino acid.
-129	Elimination of the immonium ion of arginine.
-156.1	Molecular weight of arginine without water; elimination of the amino acid.
-175	Molecular weight of arginine including water and one hydrogen; full elimination, leading to a radical ion.

For all spectra and all diagnostic losses the relative abundances of the windows (normalized to the total ion current) were recorded and then submitted to data mining. Since the charge state was known, the diagnostic fragments for doubly charged precursors were divided by two and for triply charged precursors by three. Although the resulting mass changes, the nomenclature is not changed here, for convenience. Due to this data mining is also simplified because the addition of parameters only valid in certain subsets of the overall data set can be avoided. Consequently, all charges are treated with the same learned model. However, many classification algorithms require two classes and cannot work on data sets with more than two classes. Therefore the three class problem (arginine, lysine, and other) was split up into two data sets. One approach investigates whether the c-terminus is arginine or any other amino acid. The other whether the c-terminus is lysine or not.

3.1. *Determining C-Terminal Arginine*

In order to investigate the importance of the selected diagnostic losses, their information content was ranked using a number of algorithms available in Orange Canvas (Table 2).

Most algorithms agree in general, except for SVM Weight. It can be gathered from Table 2 that the first six parameters contain the majority of the information needed to distinguish between arginine and other amino acids at the c-terminus. Nonetheless, all parameters were included when training the learners since they may still contribute some of the distinguishing power.

When looking at the resulting decision tree (Figure 1) which was created with a different algorithm from the ones in Table 2, some of the parameters, that did not receive good scores in Table 2, are used early on in the decision trees which means they have great distinguishing power between classes.

The collected data was analyzed using the supervised learning algorithms available in Orange Canvas. As expected, not all algorithms performed equally well, with the best one being Random Forest (Table 3).The distance to k-Nearest Neighbor, CN2 Rules, and Classification Tree are not very significant, though. The following algorithm, Logistic Regression

Table 2. The information content of the attributes according to different measures is presented. Rows are sorted in respect to Random Forest since its classification was most accurate for predicting the arginine or other status of peptide c-termini.

Diagnostic Loss	Random Forest	ReliefF	Inf. Gain	Gain Ratio	Gini	SVM Weight
-156.1	6.57	0.06	0.07	0.04	0.02	36.71
-175	4.02	0.00	0.01	0.00	0.00	0.09
-42	3.20	-0.01	0.19	0.11	0.05	216.31
-43	2.67	-0.01	0.09	0.05	0.03	22.68
-17	2.34	0.04	0.12	0.06	0.03	303.88
-16	2.04	0.01	0.14	0.08	0.04	162.40
-34	1.65	0.00	0.15	0.08	0.04	277.53
-33	1.36	0.01	0.07	0.04	0.02	89.99
-129	1.28	0.00	0.05	0.04	0.02	132.19
-57	0.88	-0.02	0.01	0.01	0.00	14.61
-128.1	0.78	0.01	0.04	0.03	0.01	43.33
-32	0.62	0.00	0.00	0.00	0.00	8.18

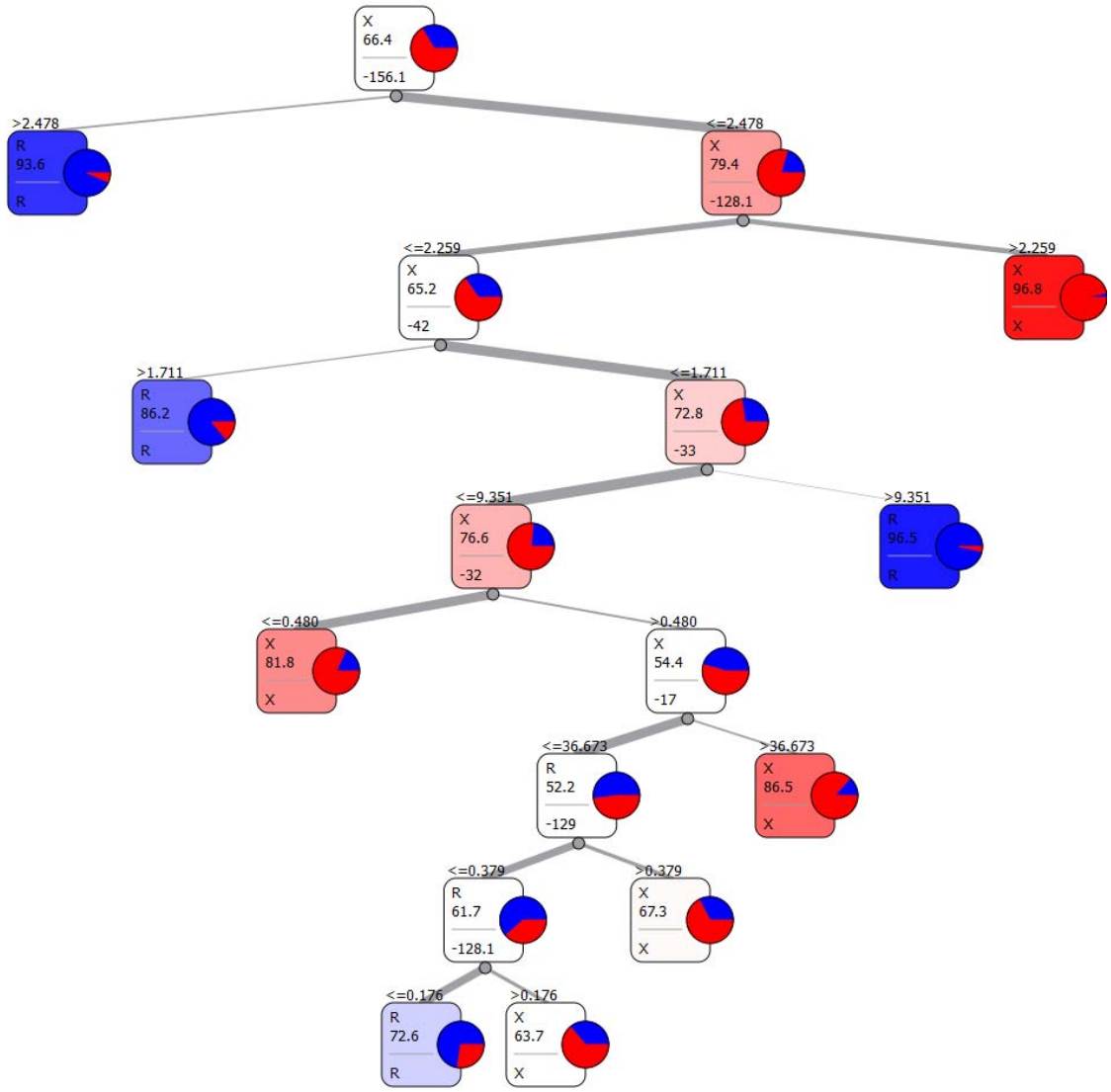

Figure 1. Decision tree for determining whether an MS/MS spectrum has a C-Terminal R. For best split exhaustive search was applied. When nodes reached more than 95% for majority class or less than 10 examples remained in a leaf no further splits were made. The nodes depict the majority class and its percentage in the upper left corner. Below the line in the nodes the next splitting parameter is shown. The pie chart graphically conveys the distribution of examples between classes. When classes become more pure the node is colored with a stronger shade of the associated color (red for any amino acid, and blue for arginine).

performs, however, much worse on this data.

Although the Random Forest algorithm achieves the best results on this data, it does not allow the construction of a decision tree within Orange Canvas, which could be implemented into software. Therefore, the Classification Tree algorithm was used to construct a decision tree (Figure 1) although it is slightly less accurate (Table 3).

Figure 1 shows that the most discriminating parameter is -156.1 which is used for the initial split of the data. This is followed by -128.1 and then by -42. This is slightly different from the data presented in Table 1 which is due to the fact that a different algorithm is employed. After the 6th split there is not much further benefit to prediction accuracy as the number of examples strongly decrease and the accuracy of the splits also decrease (Figure 1). This analysis is only for the decision whether a peptide precursor terminates with arginine or not. Additionally, the decision between lysine and other amino acids needs to be made.

3.2. *Determining C-Terminal Lysine*

The same process made for predicting arginine as the terminal amino acid was repeated for lysine. First the information content is ranked according to several algorithms. The assumption here is that different parameters should be important for the decision than the ones observed for arginine. Obviously, there is some overlap of important parameters such as -156.1 and -42 (Table 4). However, whether the

Table 3. Classification accuracy and other quality measured for all classification algorithms available in Orange Canvas for the prediction of terminal arginine or any other amino acid. The table is ordered by classification accuracy.

Classification Algorithm	Accuracy	Sensitivity	Specificity	AUC	F1	Precision	Recall
Random Forest	0.9022	0.8942	0.9157	0.9684	0.9200	0.9473	0.8942
kNN	0.9015	0.9252	0.8613	0.9599	0.9220	0.9188	0.9252
CN2 rules	0.8860	0.9390	0.7962	0.9341	0.9120	0.8865	0.9390
Classification Tree	0.8696	0.8988	0.8201	0.9378	0.8966	0.8944	0.8988
Logistic Regression	0.7758	0.7602	0.8022	0.8655	0.8101	0.8669	0.7602
Naive Bayes	0.7416	0.6618	0.8769	0.8408	0.7632	0.9011	0.6618
SVM	0.7281	0.6647	0.8356	0.8328	0.7546	0.8727	0.6647
Majority	0.6290	1.0000	0.0000	0.5000	0.7722	0.6290	1.0000

absence or presence of a parameter is important is not conveyed in this analysis. Thus the absence of a parameter could be support in one case whereas in the other case its presence could mean support. It can be observed that the most important parameter in Table 4 is one of the least important ones in Table 2.

So the assumption holds and different parameters or a different semantic of the same parameter are important for lysine prediction. This difference is further supported by the significantly different make of the decision tree (Figure 2) when compared to the arginine decision tree (Figure 1).

Just like for the classification of terminal arginine, several supervised learning algorithms were tried for learning from the data (Table 5). In this case the k-Nearest Neighbor (kNN) algorithm separates best among the two classes (lysine and other). In this analysis, the distance between the top scoring algorithms is even closer than for the prediction of arginine. Only Majority vote is significantly worse on this data than the other algorithms.

The classification accuracy for decision between lysine and other amino acids is slightly better than for the decision between arginine and other amino acids. Unfortunately, kNN does not produce rules or decision trees, either so that again the Classification Tree algorithm provided by Orange Canvas had to be employed for visualizing how decisions can be made between the classes (Figure 2).

Figure 2 shows that the decision for whether the c-terminus is lysine or any other amino acid is much more involved than for arginine (see Figure 1). Another aim for making the decision tree was to implement it into a rule based system for predicting the c-terminal amino acid of the precursor of an MS/MS spectrum.

Unfortunately, it turned out that this is only possible if the charge state of the precursor is known. Furthermore, there is accuracy dependence between charge and observed accuracy with higher charges leading to less accurate classification (Table 6). Therefore, software was designed to perform different decisions, if the charge is known and default to charge one if it is not provided.

3.3. *Software Implementation of the RKDecider*

The data mining results presented above show the theoretically possible accuracy based on the measured data set. One complication that would occur in practice is that the charge may not be annotated for recorded mass spectra. In order to develop software, which can be used to predict the status of the c-terminal amino acid (arginine, lysine, or other) the data set was split into six subsets based on the charge state and the terminal amino acid. For each resulting subset, a decision tree was built using Orange Canvas. These decision trees were subsequently implemented into software, named RKDecider.

This software, available as a console application, that implements the decision trees produced by Orange Canvas and combines them into one algorithm is available for download at: http://bioinformatics.iyte.edu.tr/RKDecider. The application first decides whether an incoming spectrum has a c-terminal arginine based on a decision tree (Supplementary File 2). The first check is for arginine as the prediction accuracy is higher than for lysine prediction (Table 6). The spectra that were not assigned arginine status are checked using the decision trees built for terminal lysine (Supplementary File 2). Finally, all unassigned spectra are labeled unknown. This either means that the spectral quality was not well enough, or did not contain the diagnostic peaks, to allow a proper identification or that the spectrum does not originate from a tryptic peptide like 90% of all protein terminal tryptic fragments which do not have a c-terminal lysine or arginine.

For this split data set, the accuracies that can be reached are limited by the algorithm which was used for decision tree building. The regular Classification Tree building algorithm in Orange Canvas was used to produce the decision trees (Supplementary File 2).

The implementation was tested on about 8500 LTQ spectra

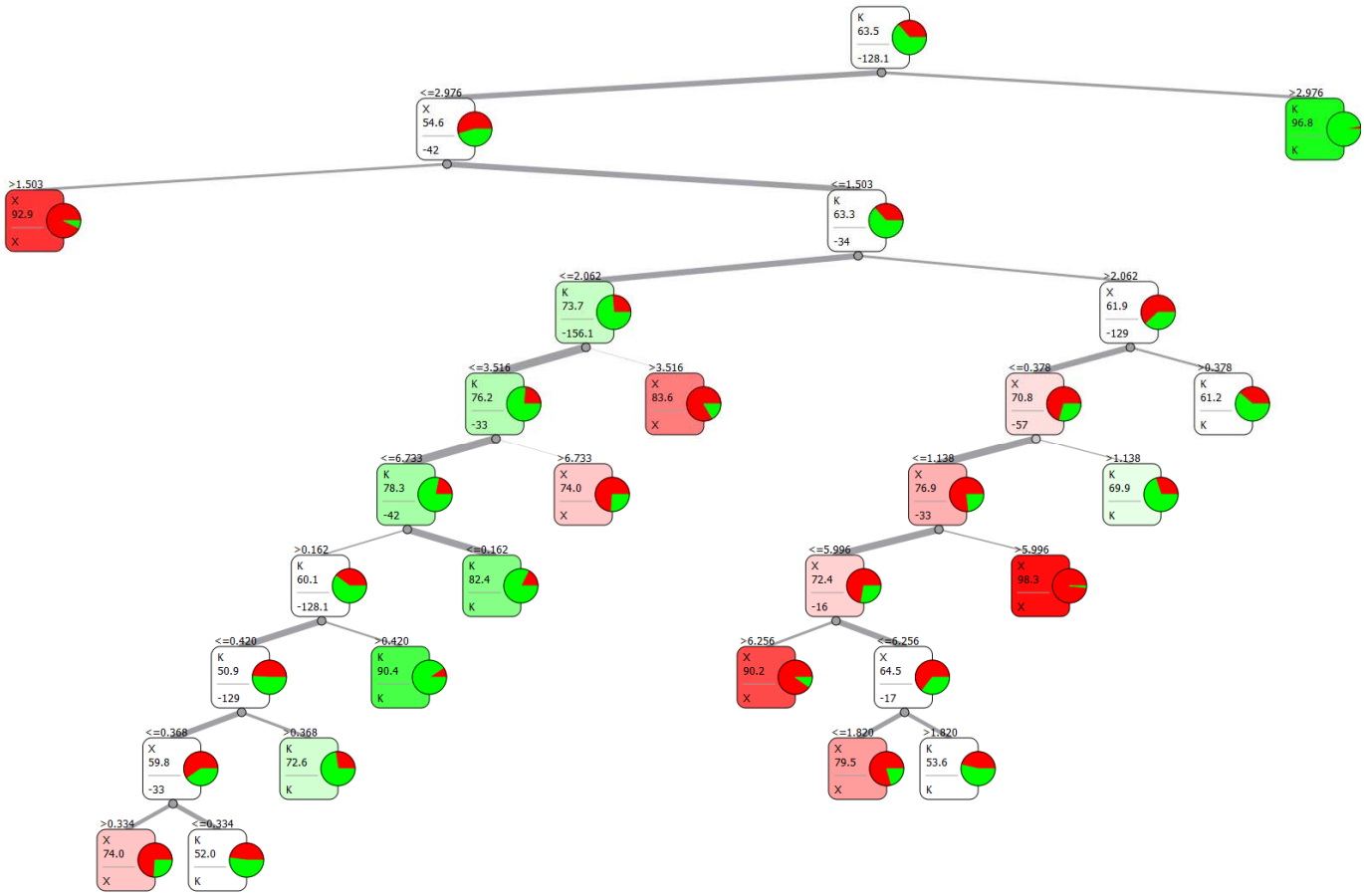

Figure 2. Decision tree for determining whether an MS/MS spectrum has a C-Terminal K. For best split exhaustive search was applied. When nodes reached more than 95% for majority class or less than 10 examples remained in a leaf no further splits were made. The nodes depict the majority class and its percentage in the upper left corner. Below the line in the nodes the next splitting parameter is shown. The pie chart graphically conveys the distribution of examples between classes. When classes become more pure the node is colored with a stronger shade of the associated color (red for any amino acid, and green for arginine).

Table 4. The information content of the attributes according to different measures is presented. Rows are sorted in respect to Random Forest since its classification was most accurate for predicting the lysine or other status of a peptide c-terminus.

Diagnostic Loss	Random Forest	ReliefF	Inf. Gain	Gain Ratio	Gini	SVM Weight
-128.1	9.30	0.03	0.22	0.12	0.06	131.42
-156.1	7.05	0.02	0.19	0.13	0.06	0.80
-34	6.18	0.01	0.20	0.10	0.06	34.51
-42	5.51	0.03	0.21	0.13	0.07	17.29
-17	3.53	0.02	0.05	0.02	0.02	1.64
-33	3.22	0.00	0.13	0.07	0.04	36.35
-16	2.32	0.00	0.11	0.06	0.04	4.72
-129	1.77	0.00	0.08	0.05	0.02	40.68
-43	1.68	0.02	0.07	0.04	0.02	0.23
-57	1.57	0.01	0.01	0.00	0.00	0.25
-175	0.89	0.00	0.05	0.03	0.02	2.26
-32	0.84	0.00	0.01	0.01	0.00	0.28

and achieved a combined accuracy of 80%. The implementation's accuracy of 80%, which in comparison to the accuracy of about 90%, achieved during data mining, seems low, but may be largely due to the fact that the spectra that were used for testing, had the annotation of their proper charge removed so that many larger peptides are likely predicted wrong. Since missing charge annotation is a general problem in mass spectrometry-based proteomics, charges were not given to the algorithm in this test.

4. Conclusion and Outlook

There is a need to know the identity of the c-terminal amino acid of a peptide. It can help to pre-filter database search candidates or to adjust database search settings so that the scope can easily be extended to include peptides with non tryptic termini, something which currently is controversial (Olsen et al., 2004). Furthermore, the confidence in the results can be elevated if the reported terminal amino acid is equal to the one predicted by RKDecider. It can be of benefit for *de novo* sequencing or sequence tag searches since it fixes one variable and thus leads to more precise results. This in-

Table 5. Classification accuracy and other quality measured for all classification algorithms available in Orange Canvas for the classification between terminal lysine and any other amino acid. The table is ordered by classification accuracy.

Classification	Accuracy	Sensitivity	Specificity	AUC	F1	Precision	Recall
Random Forest	0.9236	0.9390	0.8969	0.9686	0.9398	0.9406	0.9390
kNN	0.9148	0.9646	0.8282	0.9455	0.9349	0.9071	0.9646
CN2 rules	0.9081	0.9278	0.8739	0.9011	0.9277	0.9275	0.9278
Classification	0.8969	0.9894	0.7361	0.9792	0.9241	0.8670	0.9894
Logistic	0.8460	0.9536	0.6590	0.9259	0.8871	0.8294	0.9536
Naive Bayes	0.8280	0.9517	0.6130	0.8991	0.8754	0.8104	0.9517
SVM	0.8241	0.9441	0.6156	0.9033	0.8720	0.8102	0.9441
Majority	0.6348	1.0000	0.0000	0.5000	0.7766	0.6348	1.0000

Table 6. Classification accuracy at different charges and for separate prediction of terminal arginine or terminal lysine. Two algorithms from the Orange Canvas package were used, classification tree and random forest.

		Classification Accuracy	
Charge	C-Terminus	Classification Tree	Random Forest
1	R vs. other	98.45	99.51
2	R vs. other	82.68	88.42
3	R vs. other	82.41	86.39
1	K vs. other	97.66	98.51
2	K vs. other	82.68	88.42
3	K vs. other	82.41	86.39

formation can also be used as a simple pre-selection filter which would just reject all peptides that do not seem to have a proper tryptic c-terminus.

About 8500 LTQ CID spectra were recorded for 44 synthetic peptides and this data was mined for rules that could predict whether the c-terminus is arginine, lysine, or some other amino acid. The overall data mining accuracy peaked at around 90%. The rules, that were learned, have been implemented into software which achieves an accuracy of 80% and is freely available at http://bioinformatics.iyte.edu.tr/RKDecider.

In the future, it will be important to extend this analysis to other mass spectrometers and fragmentation methods, but for this to be successful, proper, accurate, and thus trustable benchmark data sets need to be created first [11]. The algorithm can also be improved by incorporating additional diagnostic losses which either support or contradict the existence of a specific amino acid at the c-terminus.

5. Supplementary material

Supplementary data and information is available at: http://www.jiomics.com/index.php/jio/rt/suppFiles/137/0

Acknowledgements

The author would like to thank Talat Yalcin and the Biological Mass Spectrometry and Proteomics facility at the Izmir Institute of Technology for their support and for vivid discussions about manuscript and project. This work was in part supported by an award for outstanding young scientists, received from the Turkish Academy of Sciences (TÜBA, GEBIP, www.tuba.gov.tr).

References

[1] M. Mann, R.C.C. Hendrickson, A. Pandey, Annual Review of Biochemistry 70 (2001) 437–73. DOI: 10.1146/annurev.biochem.70.1.437

[2] S.P. Mirza, M. Olivier, Physiological Genomics 33 (2008) 3–11. DOI: 10.1152/physiolgenomics.00292.2007

[3] H. Steen, M. Mann, Nat Rev Mol Cell Biol 5 (2004) 699–711.

[4] J. V Olsen, S.-E. Ong, M. Mann, Molecular & Cellular Proteomics: MCP 3 (2004) 608–14. DOI: 10.1074/mcp.T400003-MCP200

[5] S. Aiche, K. Reinert, C. Schütte, D. Hildebrand, H. Schlüter, T.O.F. Conrad, PloS One 7 (2012) e40656. DOI: 10.1371/

journal.pone.0040656

[6] E. Kapp, F. Schütz, Current Protocols in Protein Science / Editorial Board, John E. Coligan ... [et Al.] Chapter 25 (2007) Unit25.2. DOI: 10.1002/0471140864.ps2502s49

[7] J. Allmer, Expert Rev. Proteomics 8 (2011) 645–57. DOI: 10.1586/epr.11.54

[8] J.M. Wells, S.A. McLuckey, Methods in Enzymology 402 (2005) 148–185. DOI: 10.1016/S0076-6879(05)02005-7

[9] F. Schütz, E.A. a Kapp, R.J.J. Simpson, T.P.P. Speed, F. Schutz, Biochemical Society Transactions 31 (2003) 1479–83. DOI: 10.1042/

[10] M. Garzotti, M. Hamdan, Rapid Communications in Mass Spectrometry: RCM 12 (1998) 843–8. DOI: 10.1002/(SICI)1097-0231(19980715)12:13<843::AID-RCM250>3.0.CO;2-W

[11] J. Allmer, Journal of Integrated OMICS (2012). DOI: 10.5584/jiomics.v2012i2012.113

[12] T. Curk, J. Demsar, Q. Xu, G. Leban, U. Petrovic, I. Bratko, G. Shaulsky, B. Zupan, Bioinformatics 21 (2005) 396–8. DOI: 10.1093/bioinformatics/bth474