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.

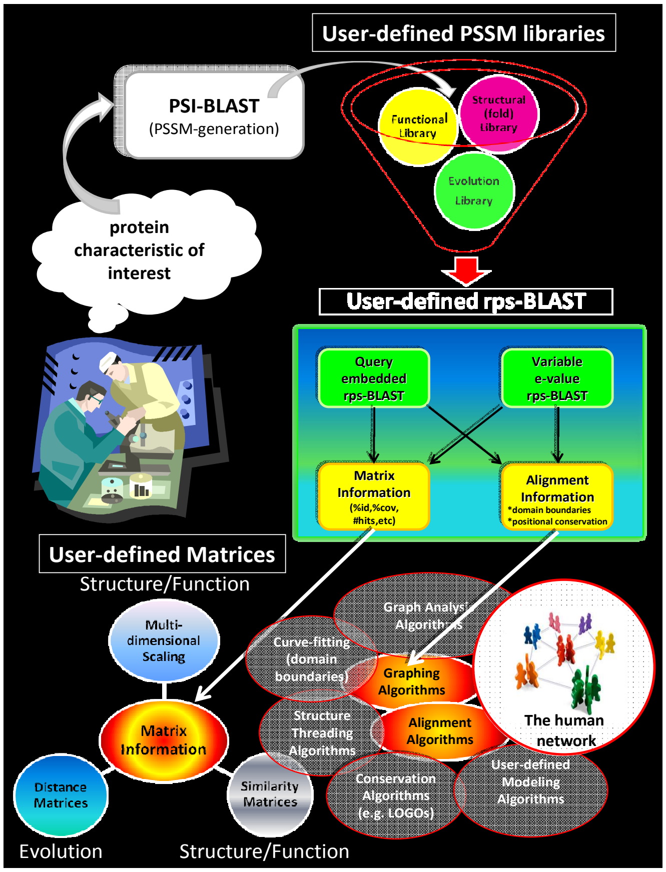

User-defined PSSM libraries
PSI-BLAST
(PSSM-generation)
Functional Library
Structural (fold) Library
Evolution Library
protein characteristic of interest
User-defined rps-BLAST
Query embedded rps-BLAST
Variable e-value rps-BLAST
Matrix Information (%id,%cov, #hits,etc)
Alignment Information *domain boundaries *positional conservation
User-defined Matrices
Structure/Function
Multi-dimensional Scaling
Matrix Information
Distance Matrices
Similarity Matrices
Evolution
Structure/Function
Graph Analysis Algorithms
Curve-fitting (domain boundaries)
Graphing Algorithms
Structure Threading Algorithms
Alignment Algorithms
Conservation Algorithms (e.g. LOGOs)
User-defined Modeling Algorithms
The human network

**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

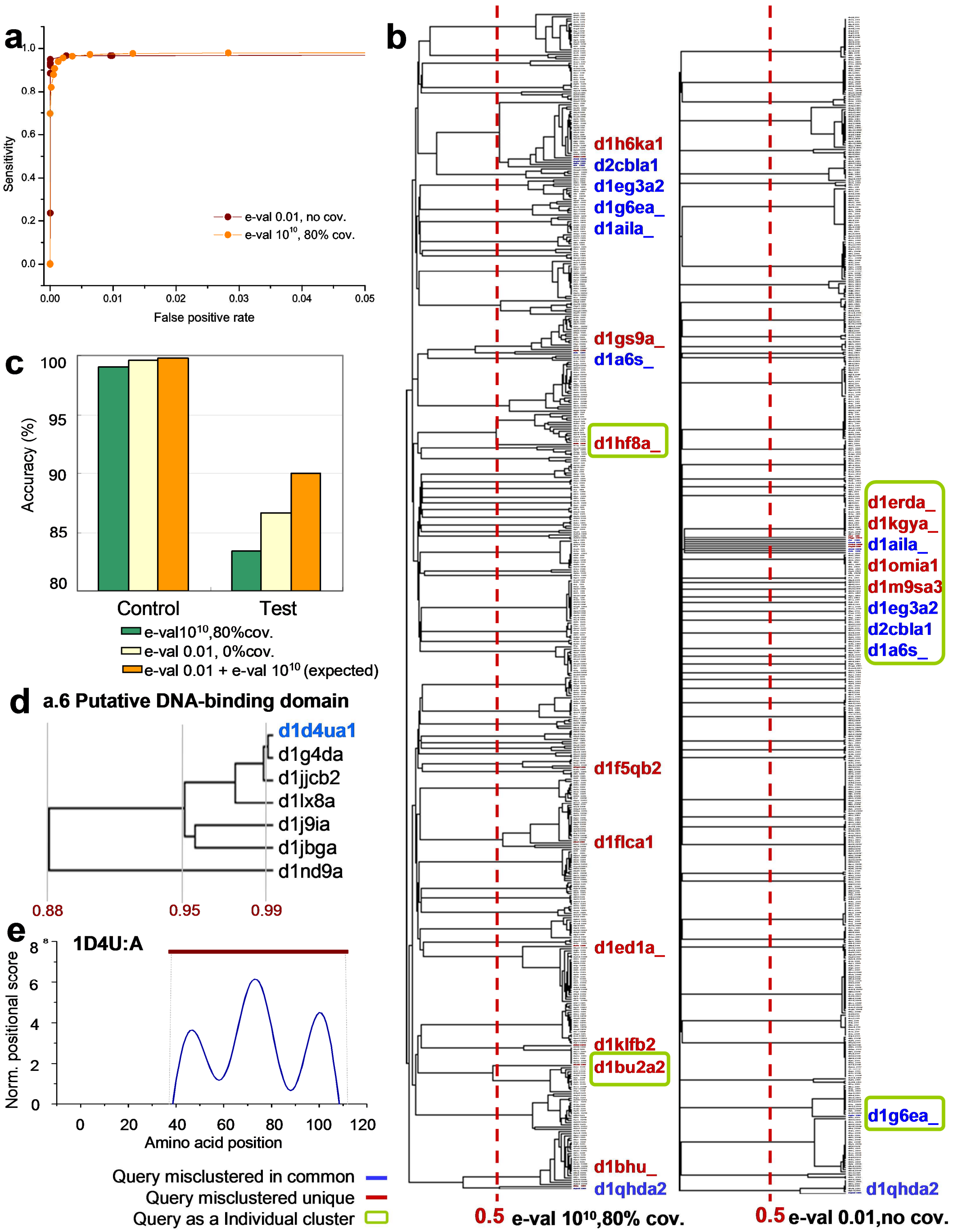


**Figure 2.** *Fold Recognition and Fold Clustering with the Fold-specific Libraries.* (a) ROC curves of Ada-BLAST. TZ-SABmark queries were encoded in a vector using rps-BLAST alignments of e-value 0.01 and 1010 with 61 fold-specific PSSM libraries (see the Supplemental Methods). (b) Hierarchical clustering 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

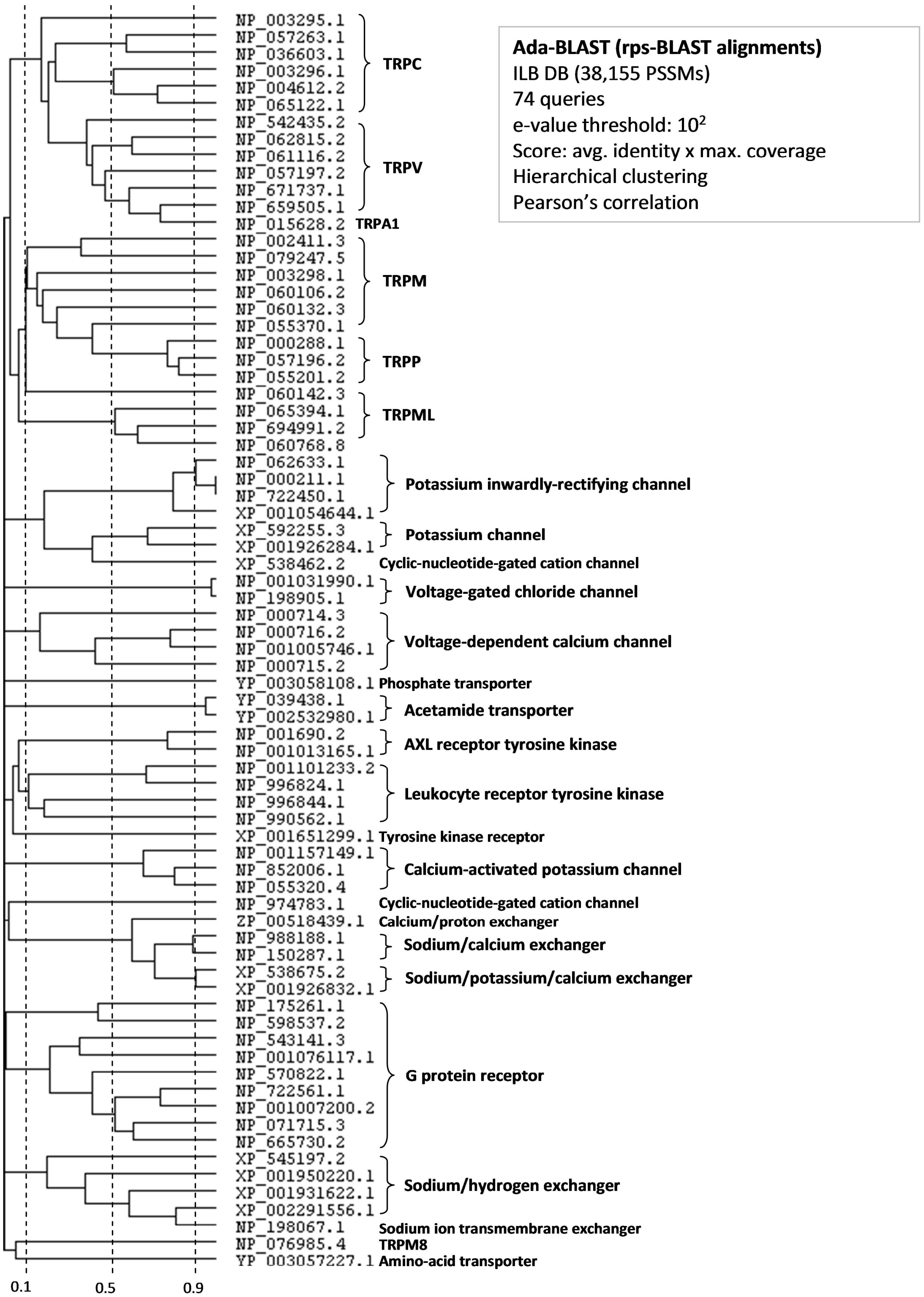


**Figure 3.** *A Classification based on Heirarchical Clustering.* 74 sequences representing multiple classes of transmembrane containing proteins were hierarchically clustered and visualized by Cluster and Treeview [34]. The dotted lines represent the correlation scores derived from the analysis. Alignments 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-

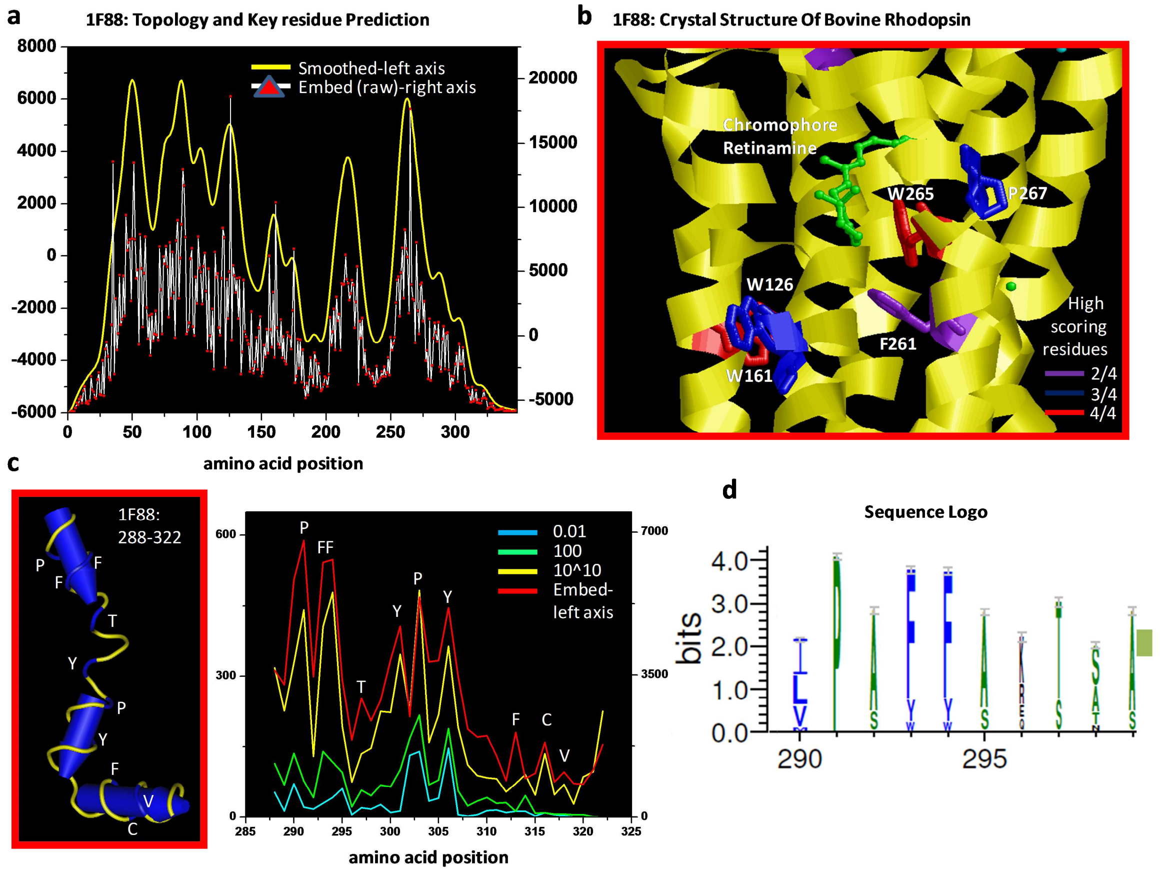


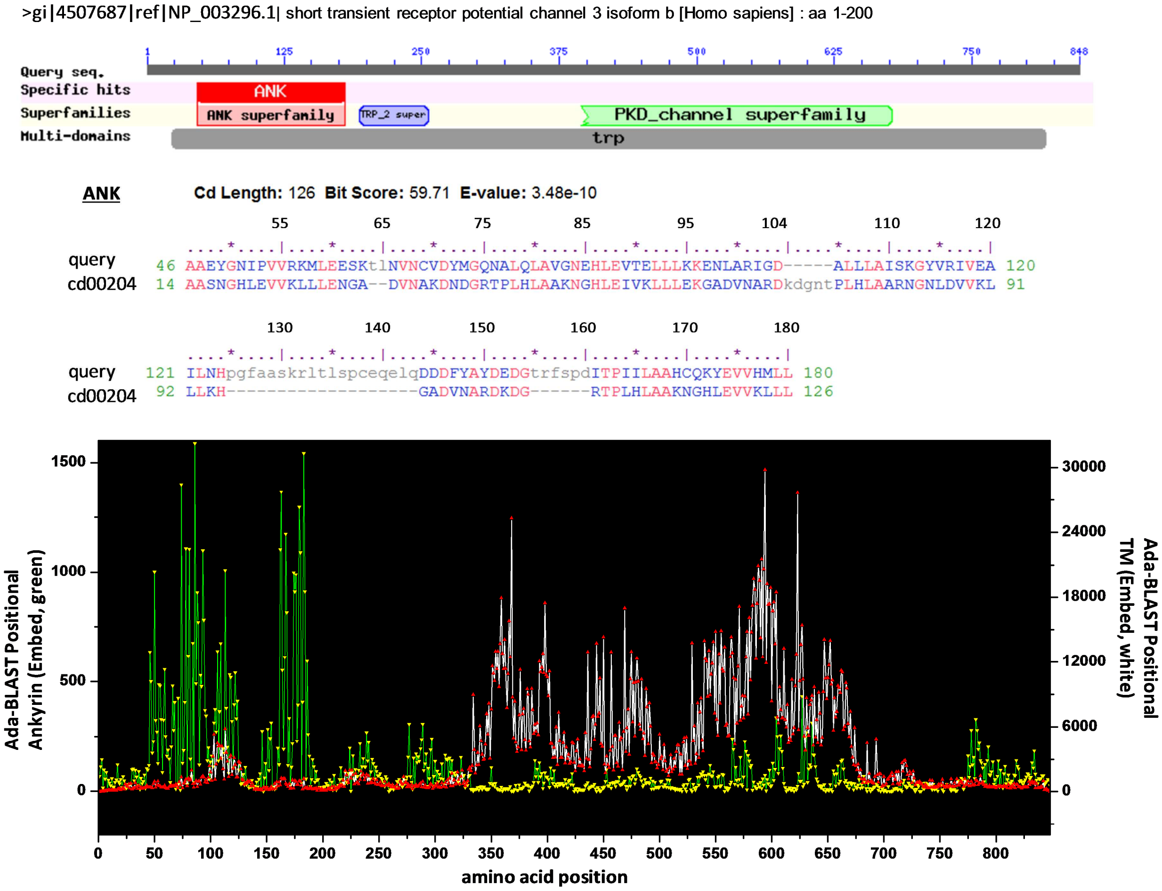
**Figure 4.** *Ada-BLAST analysis of Bovine Rhodopsin.* (a) The primary amino acid sequence of 1F88 was run against ILB-DB in Ada-BLAST (embedding setting). This graph depicts normalized raw data (right axis) compared with the data smoothed by Fast Fourier Transform (setting=8). These data were normalized 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-



>gi|4507687|ref|NP_003296.1| short transient receptor potential channel 3 isoform b [Homo sapiens] : aa 1-200

```
ANK        Cd Length: 126  Bit Score: 59.71  E-value: 3.48e-10

                      55        65        75        85        95        104           110       120
                ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
query      46  AAEYGNIPVVRKMLEESKtlNVNCVDYMGQNALQLAVGNEHLEVTELLLKKENLARIGD-----ALLLAISKGYVRIVEA  120
cd00204    14  AASNGHLEVVKLLLENGA--DVNAKDNDGRTPLHLAAKNGHLEIVKLLLEKGADVNARDkdgntPLHLAARNGNLDVVKL  91

                      130       140       150       160       170       180
                ....*....|....*....|....*....|....*....|....*....|....*....|
query     121  ILNHpgfaaskrltlspceqelqDDDFYAYDEDGtrfspdITPIILAAHCQKYEVVHMLL  180
cd00204    92  LLKH-------------------GADVNARDKDG------RTPLHLAAKNGHLEVVKLLL  126
```

**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

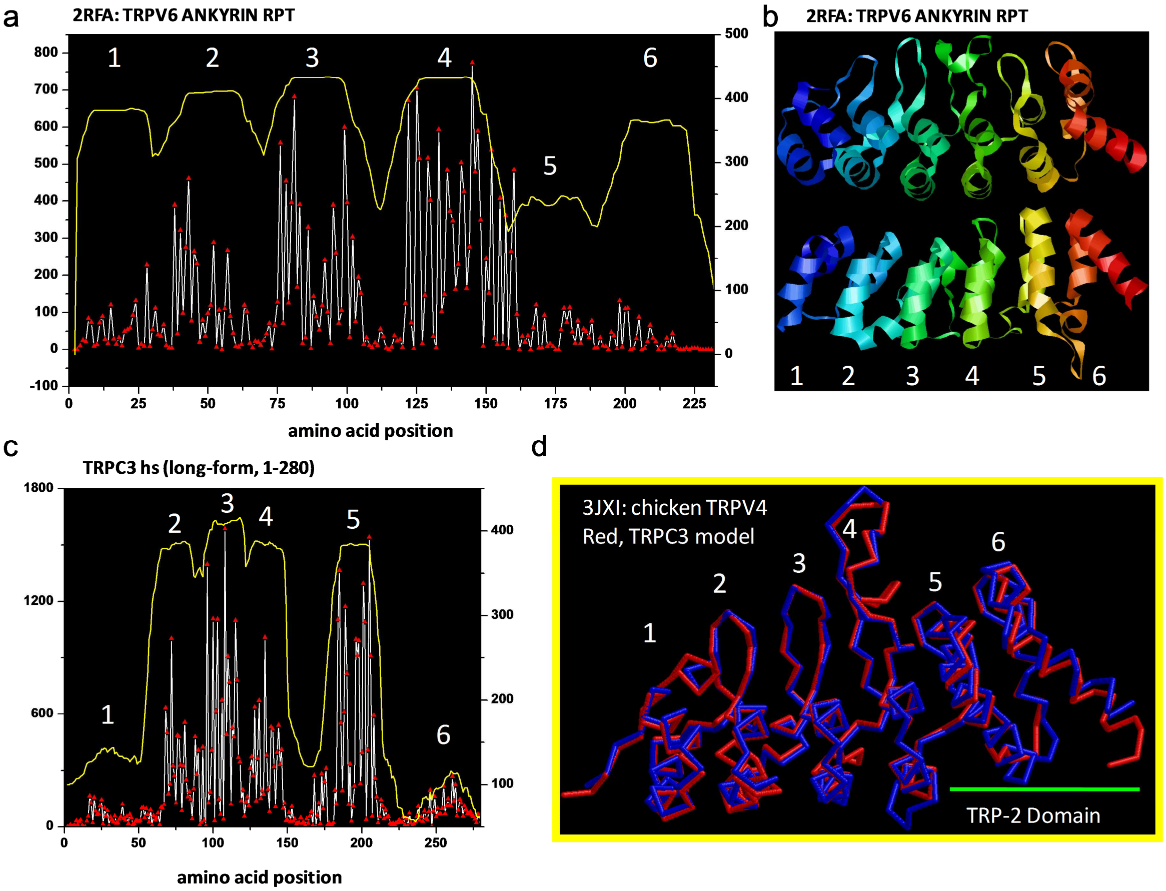


**Figure 6.** *Anykyrin repeat modeling of TRPC3.* (a) Ada-BLAST ANK ILB measurements (embedded rps-BLAST) of the structurally resolved ankyrin repeats in TRPV6 (*overlapping alignment output*- yellow line, *amino-acid conservation output*- white,red). (b) Two views of the structure of TRPV6 with each 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

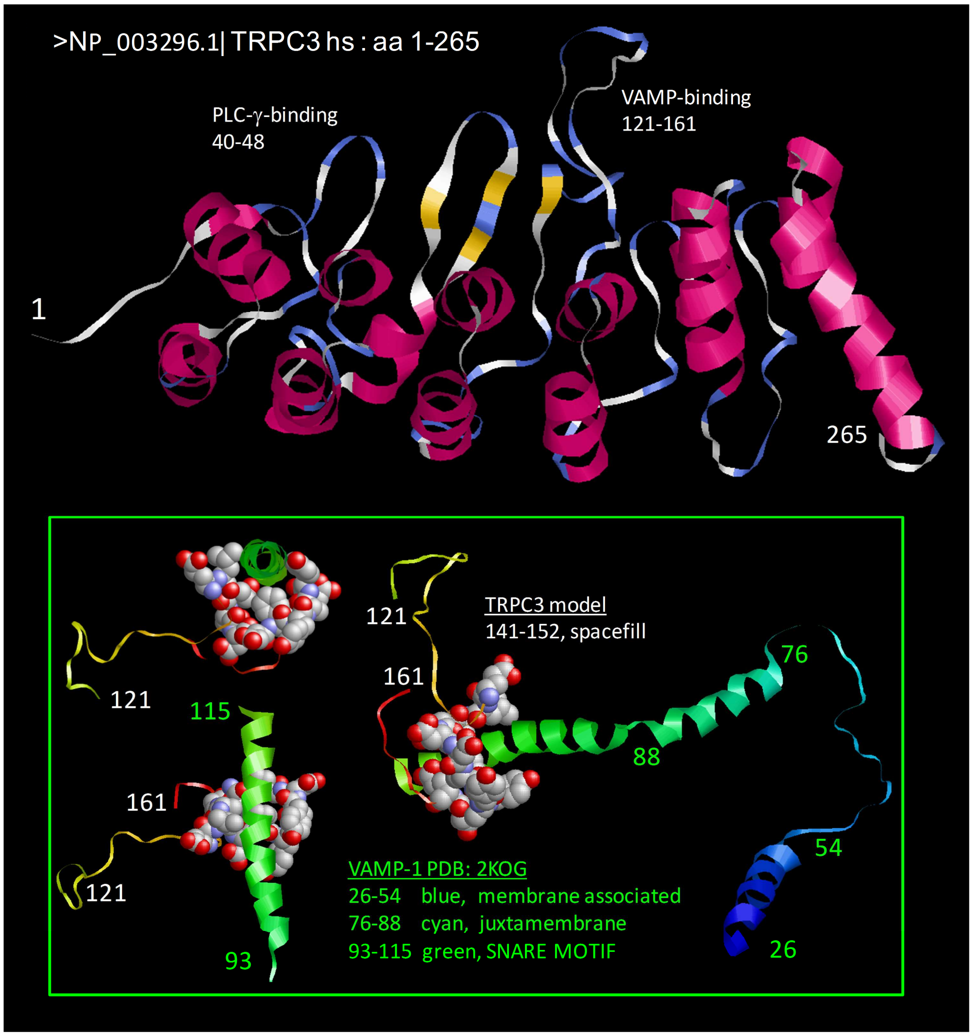

>NP_003296.1| TRPC3 hs : aa 1-265
PLC-γ-binding
40-48
VAMP-binding
121-161
1
265
121
115
161
121
93
121
161
TRPC3 model
141-152, spacefill
76
88
54
26
VAMP-1 PDB: 2KOG
26-54 blue, membrane associated
76-88 cyan, juxtamembrane
93-115 green, SNARE MOTIF

**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-

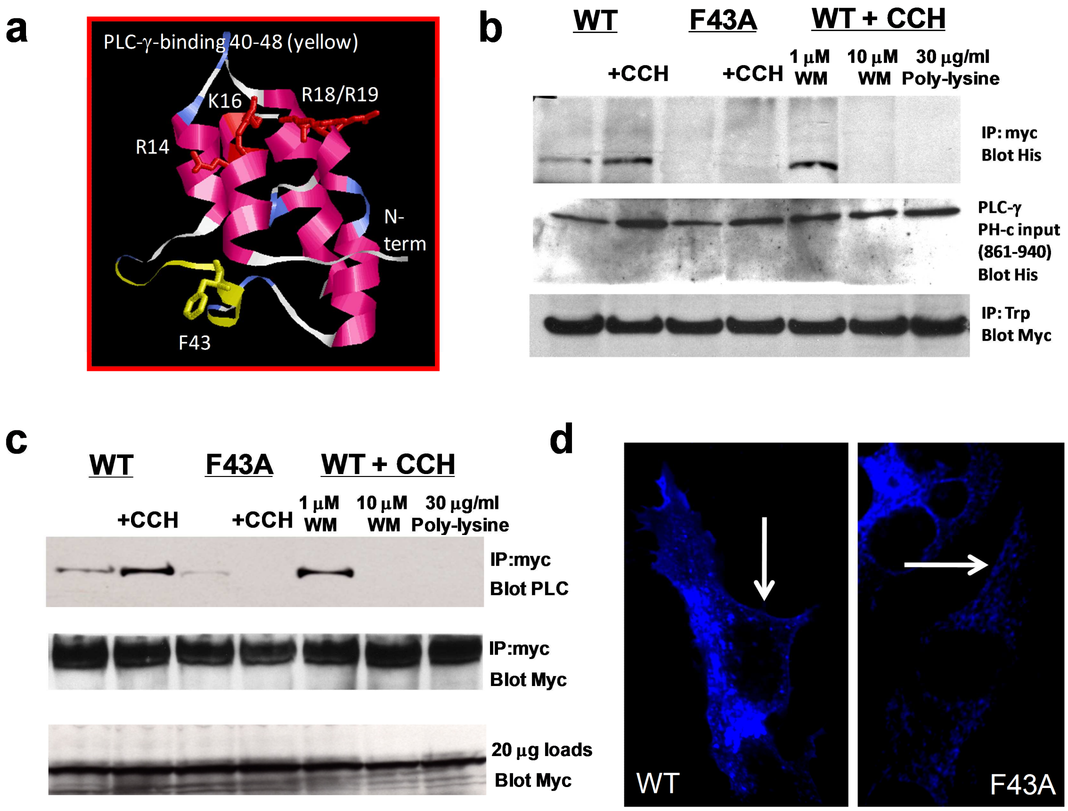


**Figure 8.** *Experimental validation of the phospholipase C-γpartial PH-domain binding to the ankyrin repeats in TRPC3.* (a) Structural model of the first two ankyrin repeats in TRPC3. Key PLCγ binding residues discovered in (ref) are labeled. (b) Co-immunoprecipitation assays of overexpressed full-length 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-

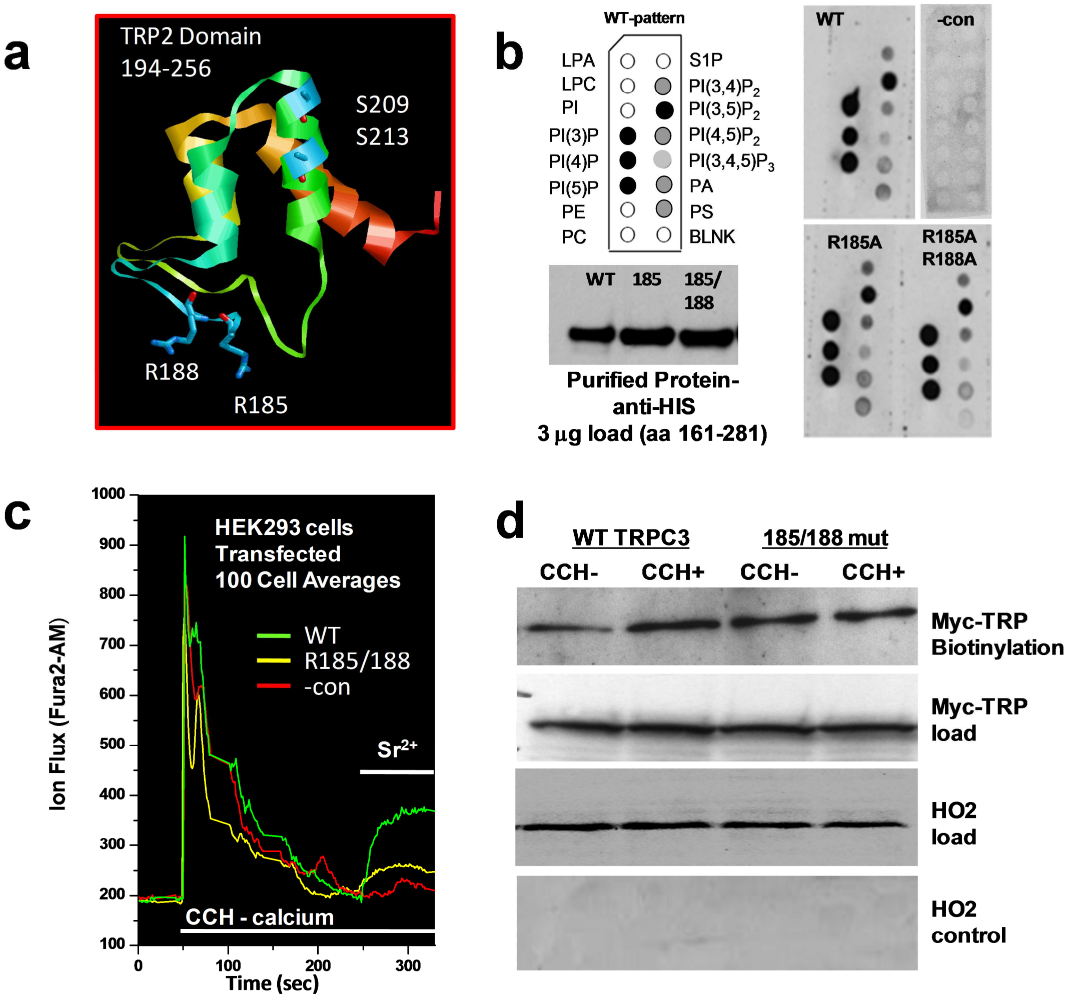


**Figure 9.** *Experimentally modeling the TRP_2 domain in TRPC3.* (a) Structural model of the first two ankyrin repeats in TRPC3. Key serines (S209, S213) involved in regulating lipid binding discovered in [23] are labeled, as well as two arginines (R185,R188) which we speculated to participate in lipid-binding 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.