BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

MOTIPS: Automated Motif Analysis for Predicting Targets of Modular Protein Domains

Hugo YK Lam[†1], Philip M Kim[*†2,8], Janine Mok[3,9], Raffi Tonikian[4,5], Sachdev S Sidhu[4,5], Benjamin E Turk[6], Michael Snyder[3,10] and Mark B Gerstein[*1,2,7]

Abstract

Background: Many protein interactions, especially those involved in signaling, involve short linear motifs consisting of 5-10 amino acid residues that interact with modular protein domains such as the SH3 binding domains and the kinase catalytic domains. One straightforward way of identifying these interactions is by scanning for matches to the motif against all the sequences in a target proteome. However, predicting domain targets by motif sequence alone without considering other genomic and structural information has been shown to be lacking in accuracy.

Results: We developed an efficient search algorithm to scan the target proteome for potential domain targets and to increase the accuracy of each hit by integrating a variety of pre-computed features, such as conservation, surface propensity, and disorder. The integration is performed using naïve Bayes and a training set of validated experiments.

Conclusions: By integrating a variety of biologically relevant features to predict domain targets, we demonstrated a notably improved prediction of modular protein domain targets. Combined with emerging high-resolution data of domain specificities, we believe that our approach can assist in the reconstruction of many signaling pathways.

Background

Important protein-protein interactions (e.g., those involved in signal transduction) are often mediated by modular protein domains [1]. These domains often work in a mix-and-match fashion, thereby acting as the building blocks of signaling pathways [2]. Examples include the SH3 and WW domains that bind proline-rich motifs [3], and the serine/threonine kinase domain that specifically phosphorylates the hydroxyl group of serine and threonine [4]. Throughout we will refer to these collectively as "domains". Since these kinds of domains play an important role in the assembly, regulatory and signaling activities of the cell [3,5,6], accurate prediction of their targets is crucial to understanding many biological pathways [7,8].

As a result, various techniques have been developed to predict domain targets and to enhance the prediction. Earlier studies have tried to use consensus sequences from phage display experiments to predict the targets of peptide-binding domains [9]. Also, a modern peptide library screening approach, which is commonly used to determine phosphorylation motifs for kinases, has shown to have high accuracy in determining domain specificity [10]. Both approaches have in common that they identify the specificity of each domain in a position-specific manner, yielding a Position Specific Scoring Matrix (PSSM; also known as Position Weight Matrix, PWM). Furthermore, many studies have demonstrated various ways to improve prediction performance using genomic information. For instance, comparative genomics and secondary structure information have been used to increase the performance of SH3 target prediction [11,12].

Nevertheless, to date the prediction of biologically relevant targets of these domains has yet to be addressed in an automated and integrated fashion. To this end, we present an automated process, which integrates comparative genomic (i.e., sequence conservation) and structural genomic (i.e., surface propensity and peptide disorder) data with traditional profile scanning method to predict

[1] Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA
[2] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA
[†] Contributed equally
Full list of author information is available at the end of the article

domain targets based on experimental screening result (e.g. peptide library screening) or their derived PSSMs. The process is fully automated and implemented as an online server. The implementation is open-source and also available for download at http://motips.gerstein-lab.org.

Results and Discussion

An Automated Pipeline Process

Our approach first converts the input data into a PSSM and then normalizes it. Secondly, it scans the target proteome by using the normalized PSSM and generates a hit list of potential domain targets. Following the motif scanning, it computes the conservation score, solvent accessibility score, and disorder score for each motif hit based on the pre-computed scores for each protein residue. It then integrates these genomic features with the motif matching scores and the number of hits per protein by naïve Bayes to predict the optimal targets based upon a validated training set. Lastly, it sorts the motif hits by their likelihood of having interaction with the domain and consolidates them into unique protein hits.

Data Conversion and Normalization

A number of experimental approaches, such as phage display and peptide library screening (see Figure 1), have been developed to identify domain binding and phosphorylation targets. However, data from different experiments result in different formats that always complicate the data analysis process. To keep the process consistent and standardized, these data are converted into PSSM followed by normalization (for supported input formats, see System Implementation and Availability).

Our approach employs two different ways to normalize the input data. The first approach is designed for signal data from experiments such as from peptide library screening. It normalizes the signal score for each amino acid at each position by the following equation:

$$Z_{ca} = \frac{S_{ca}}{\sum_{i}^{m} S_{ci}} \times m \quad (1)$$

where Z_{ca} is the normalized score for amino acid a at position c, which has a signal score S_{ca}, and m is the total number of amino acids. Equation (1) thus computes the weight for each amino acid at each position and scales it up by the total number of amino acids. However, to consider the known specificity for domains such as the serine/threonine kinase domain, which have fixed amino acid targets (e.g., serine and threonine) at a certain position in the binding motif, a score of 0 is automatically assigned to every other amino acid that is not expected at that position. To indicate the slight probability of observing the fixed amino acids at other positions, a pseudo-count of 1 is assigned to each of them at these non-specific positions.

The second way of normalization is designed for peptide data from experiments such as from phage display experiment. Our approach employs the pseudo-count method based on substitution probabilities to complement the incomplete or imperfect representation of a position in the original peptide data [13]. Pseudo-counts are needed since this kind of experiments significantly undersample sequence space, thereby severely penalizing rare residues. It calculates the probability p_{ca} of amino acid a at position c by equation (2) as follows:

$$p_{ca} = \frac{n_{ca} + b_{ca}}{N_c + B_c} \quad (2)$$

$$B_c = \psi \times R_c \quad (3)$$

where n_{ca} and b_{ca} are the count and pseudo-count for amino acid a at position c, while N_c and B_c are the total

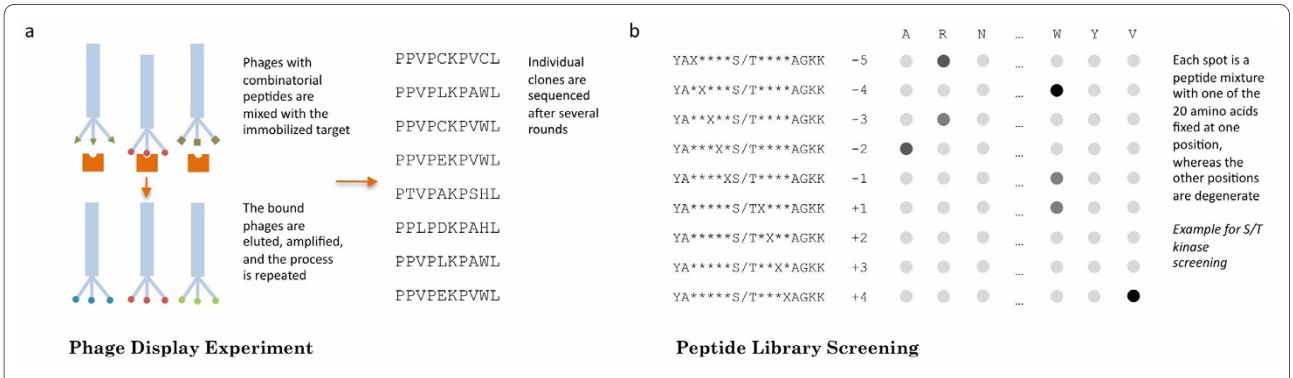

Figure 1 Experiments for Motif Identification. a) The phage display experiment identifies potential target peptides of short sequences, and b) the peptide library screening measures the binding specificity at position level. The resulting experimental data of such experiments can be converted into a Position Specific Scoring Matrix (PSSM).

count and pseudo-count for all amino acids. The total pseudo-count B_c is calculated from equation (3) with ψ as an empirically chosen positive number (default to 5) and R_c as the unique count for all amino acids at position c. Taking different substitution probabilities of different amino acids into consideration, substitution matrixes such as the BLOSUM 62 [14,15] and McLachlan [16] matrixes are used to calculate pseudo-count b_{ca} by equation (4) shown as the following:

$$b_{ca} = B_c \times \sum_i^m \frac{n_{ci}}{N_c} \times \frac{q_{ia}}{Q_i} : Q_i = \sum_i^m q_{ia} \quad (4)$$

where q_{ia} is the substitution probability for amino acid a replaced by i, and Q_i is the substitution probability for a replaced by any amino acid. In addition to the pseudo-count method based on substitution probabilities, we also provide alternative pseudo-count methods based on flat counting (adding 1 to all values) and entropy (adding a pseudo-count proportional to the entropy of each position to its corresponding values).

Motif Scanning and Scoring

To scan the target proteome for potential domain targets and to score them, our approach uses a window-sliding method based on a normalized PSSM similar to the method used in Scansite [17,18]. For each protein in the target proteome, it slides a window of size equivalent to the length of the motif on the peptide sequence by every single amino acid (see Figure 2). Based on the scoring matrix, the score for each window sequence is calculated by equation (5):

$$E' = \sum_c^l -\log_2 \left(\frac{S_{ca}}{\sum_i^m S_{ci}} \right) \quad (5)$$

where l is the length of the motif and S_{ca} is the score for amino acid a at position c in the window sequence. This equation is also used to calculate an optimal score of the motif where S_{ca} is the maximum score at position c in the scoring matrix. Then the final normalized score E for the window sequence is calculated by equation (6):

$$E = \frac{E'sequence - E'optimal}{E'optimal} \quad (6)$$

To improve the efficiency of the scanning algorithm, each motif hit is compared immediately to a sorted hit list of fixed size (currently 2,000 hits) and will only be retained if it has a more significant score than the least significant one in the list.

Structural Features and Scoring

Although a profile-matching scan could identify possible domain targets, it does not take into account the structural information of the target sequences that are also related to protein-protein interactions. For instances, sequences exposed on the surface should be more accessible than those that are buried; sequences that are

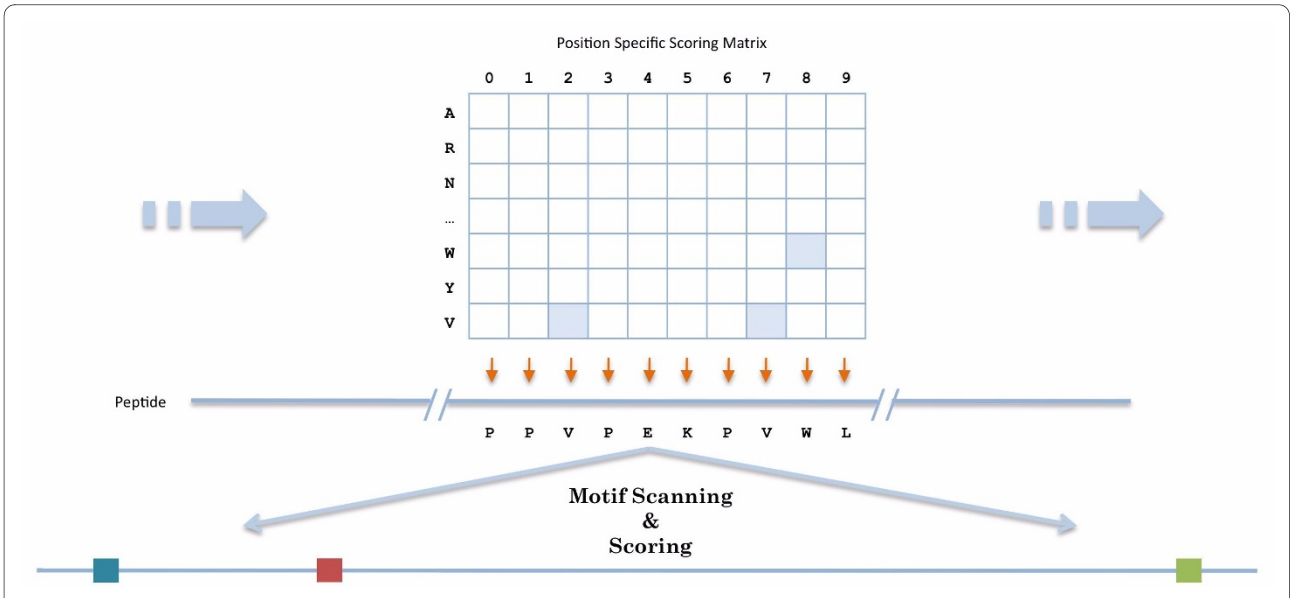

Figure 2 Motif Scanning and Scoring. Identify potential target sites of the domain by sliding a Position Specific Scoring Matrix (PSSM) across the peptides in the proteome and comparing the motif matching scores for each window.

unfolded should be more easily bound than those that are folded; and structures that are highly conserved among close species could have more biological significance. Taking these factors into account, our approach includes three major structural and conservation features in the prediction, which are surface propensity, protein disorder, and sequence conservation, to complement the motif scanning score (see Figure 3).

The degree of surface propensity of a given sequence is measured by its relative solvent accessibility, which represents the extent of residue solvent exposure. It is predicted by a protein structure prediction program, SABLE, which uses a neural network-based regression algorithm [19]. To measure the disorder of the sequence, DISOPRED, a neural networks and PSI-BLAST-based approach is used to estimate the probability of the region being disordered [20,21]. For measuring the conservation of the sequence structure, orthologs of the sequence are identified using INPARANOID [22]. Following the ortholog identification, the sequences in the orthologous groups are aligned with MUSCLE [23] and a conservation score for each position in the sequence is estimated by its entropy using AL2CO [24].

For each protein in each proteome being studied, the solvent accessibility, disorder and conservation scores are pre-computed for each residue. As a result, the scores for the motif hits could be calculated in a timely manner.

Feature Integration and Target Prediction

In addition to calculating the structural and conservation scores for each motif hit, the number of hits per protein is also calculated as a feature for the hit. Our approach then applies a Bayesian learning algorithm to integrate all the aforementioned features, including the motif scanning score, solvent accessibility score, disorder score, conservation score, and number of hits per protein, to predict potential domain targets. Because of the simplicity and efficiency of the naïve Bayes model, it is employed to build a classifier based on a validated training set under the assumption of independence of the features. In particular, the default models (i.e., the SH3 model based on Sho1 and the S/T kinase model based on Prk1) used a number of experimentally determined interaction pairs [25,26] as the gold-standard positives to train the algorithm. Moreover, a set of paired proteins in which each pair was annotated to always localize to two different compartments (for example, nucleus only and cytoplasm only in the Gene Ontology) in the cell was selected as the gold-standard negatives. The conditional probability can then be calculated from the given features based on equation (7):

$$p(I \mid F_1, \ldots, F_n) \propto p(I) \prod_{i=1}^{n} p(F_i \mid I) \tag{7}$$

where I is the class variable (i.e., interaction or non-interaction), F is the feature such as the motif scanning score, and n is the total number of features. To assess the independence of the features, pair-wise correlation coefficients were calculated. The results showed the pair-wise correlation coefficients have an average of 0.23 for the SH3 model and 0.18 for the S/T kinase model, indicating the features are to a large extent independent. Further-

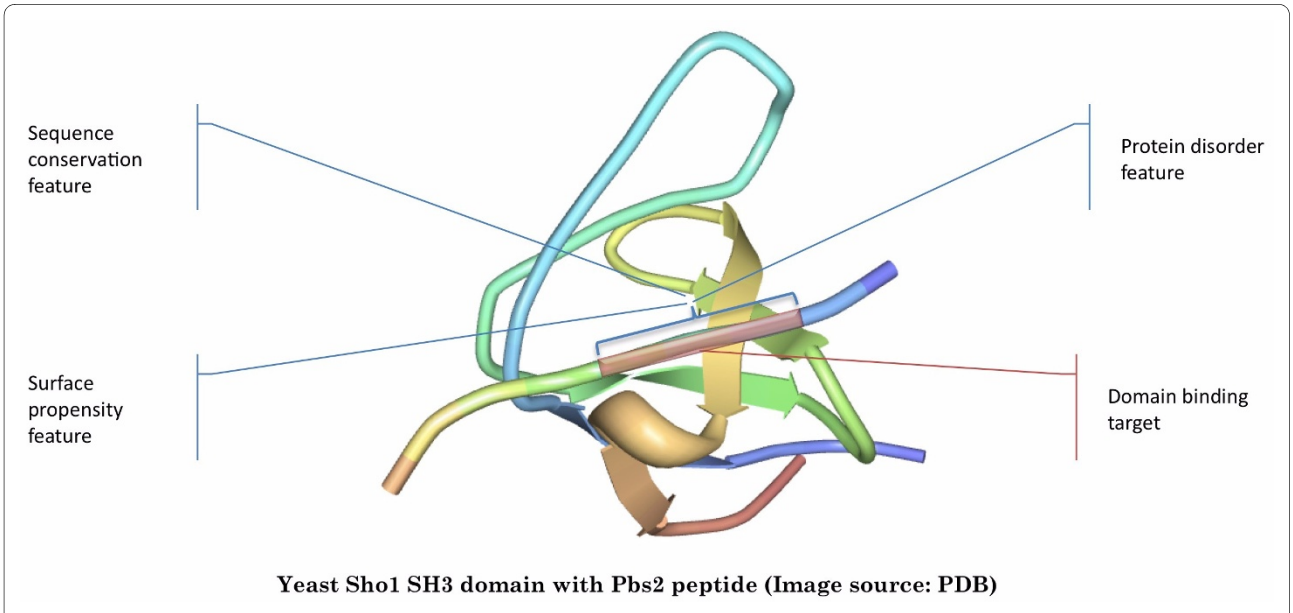

Figure 3 A Peptide-Binding Domain Example. A peptide-binding domain, such as the SH3 domain, recognizes the binding site on a peptide which exhibits certain structural and conservation features including surface propensity, protein disorder, and sequence conservation.

more, since the independency assumption is not harmful for data pre-processed with Principal Component Analysis (PCA) [27], we performed PCA to transform the possibly correlated features into uncorrelated features. The first three principal components were chosen to build a naïve Bayes model followed by a stratified 10-fold cross-validation. The Area Under Curve (AUC; 89.1 for the SH3 model and 75.9% for the S/T kinase model) of the Receiver Operating Curve (ROC) resulting from the PCA transformation was then compared to the AUC (91.8% for the SH3 model and 78.6% for the S/T kinase model) without PCA. No significant deviation of performance was observed between the predictions without PCA and those with PCA, indicating no strong dependency among the original features.

Finally, the motif hits from the domain of interest are classified under the selected model and sorted by their likelihood of having an interaction with the domain. Hits for the same protein are consolidated into one single hit represented by the most likely target. Genomic information that is not used in the prediction, such as protein-protein interaction data, localization data and phosphorylome data, could also be integrated easily with the tab-delimited hit list for further analysis while phosphorylation prediction data from mass spectrometry experiments can be used as cross-validation.

Prediction Performance

To assess the prediction performance of our approach, we benchmarked with two existing methods: 1. the Eukaryotic Linear Motif (ELM) database [28], which predicts functional sites in eukaryotic proteins by patterns with context-based rules and logical filters such as the structure filter; and 2. the Scansite method [17], which uses a motif profile-scoring approach to predict sites within proteins that are likely to be phosphorylated or bind to domains. Based on the SH3 interactome data [25], a model for the SH3 domain was trained with the Sho1 interactions. Then, we performed our prediction, requiring a likelihood value above 0.9, on 10 other different SH3 proteins by using the aforementioned model. We compared our results with the predictions from the ELM database (data retrieved from the web server using a Python program for 5 different SH3 ligands available on the server) and from the Scansite scanning (which requires a score not more than 3 fold of the optimal score). Our results (see Figure 4) show that on average our prediction has a 49% increase in accuracy in predicting the validated targets of the SH3 proteins when compared to the ELM prediction. When compared to the profile-scoring method of Scansite, our prediction is almost twice as accurate (90% higher). In addition to predicting SH3 targets, our approach was employed to predict Prk1 phosphorylation sites [26]. A stratified 10-fold cross-validation has shown a performance increase (see Figure 4; 79% AUC in a ROC curve) when compared to the profile-scoring method (72% AUC).

System Implementation and Availability

The motif analyzing process mentioned above is implemented as an online server, which allows researchers to upload their experimental data representing the motifs of the domains and to predict the targets. Our pipeline supports various input data formats. For specific analysis software, it currently supports the Gene Pix Result format http://www.moleculardevices.com/pages/software/gn_genepix_file_formats.html#gpr that is usually used for peptide library screening data, and the BRAIN project's peptide format http://www.baderlab.com/Software/BRAIN/PeptideFile that is usually used for phage display experiments. For general purposes, it supports the FASTA format (i.e., a set of peptides with the same length that represent the possible interacting sites) and the Nx20 format (i.e., a tab-delimited format that represents the positional scores of a motif profile with the first row labeled with the amino acid residues and the subsequent rows as the different positions). The pipeline currently has a compilation of 20 proteomes consisting of 14 yeast proteomes (*S. cerevisiae*, *C. albicans*, *D. hansenii*, *C. glabrata*, *K. lactis*, *N. crassa*, *S. bayanus*, *S. castelli*, *S. kluyveri*, *S. kudriavzevii*, *S. mikatae*, *S. paradoxus*, *S. pombe*, *Y. lipolytica*), 2 worm proteomes (*C. briggsae*, *C. elegans*), and 4 mammalian proteomes (*C. familiaris*, *P. troglodytes*, *M. musculus*, *H. sapiens*).

The feature scores were pre-computed and the default prediction models, which could be replaced by a user-defined training set (a tab-delimited file with the gene on the first column and a logical value on the second indicating the interaction), were also built. Moreover, the analyzing process is implemented as an asynchronous multi-threading pipeline process so the prediction results can be delivered to the users via email offline, in addition to being displayed online. Furthermore, the entire system is built using the Java programming language under a Model View Controller architecture in which the analysis process is implemented as a standalone open-sourced program. Therefore, the process could be customized by researchers and executed in command line on multiple platforms. The naïve Bayes classification is performed using Weka, the open-source Java data mining software [29].

The standalone pipeline and database are available for download at the MOTIPS server at http://motips.gersteinlab.org.

Conclusions

By integrating a variety of biologically relevant features and using a Bayesian learning algorithm to predict

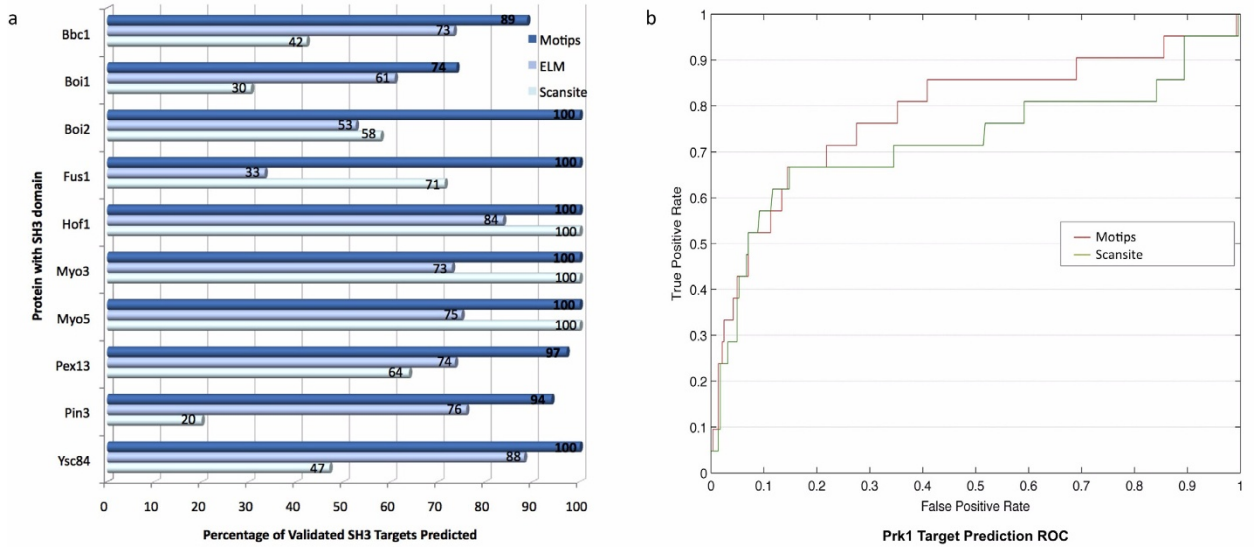

Figure 4 Targets Prediction Performance. a) The benchmark of SH3 target prediction based on the validated targets for 10 different SH3 proteins, and b) the Receiver Operating Curve (ROC) comparing the prediction performance for the Prk1 kinase targets.

domain targets, our approach has improved the domain binding and phosphorylation target predictions notably compared to using only profile-matching scan. We believe our approach is versatile enough to predict targets of domains of different kinds, and its implementation as an online public server could facilitate researchers in predicting domain targets more accurately.

Authors' contributions

HL and PK designed the methodology and drafted the manuscript. HL implemented the methodology. JM carried out the kinase specificity experiments and participated in its analysis. RT, SS, BT, MS and MG guided the study and helped to draft the manuscript. PK and MG conceived of the study. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge support from the NIH and from the AL Williams Professorship funds. We would also like to thank Chong Shou for proofreading the manuscript and Kevin Yip for the discussion on the PCA.

Author Details

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA, [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA, [3]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06520, USA, [4]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, M5S 1A8, Canada, [5]Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, M5G 1L6, Canada, [6]Department of Pharmacology, Yale University, New Haven, CT 06520, USA, [7]Department of Computer Science, Yale University, New Haven, CT 06520, USA, [8]Current Address: Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, M5S 3E1, Canada, [9]Current Address: Stanford Genome Technology Center, Department of Biochemistry, Stanford University, Palo Alto, CA 94304, USA and [10]Current Address: Department of Genetics, Stanford University, Palo Alto, CA 94305, USA

Received: 3 February 2010 Accepted: 11 May 2010
Published: 11 May 2010

References

1. Zarrinpar A, Park SH, Lim WA: **Optimization of specificity in a cellular protein interaction network by negative selection.** *Nature* 2003, **426(6967):**676-80.
2. Pawson T, Nash P: **Assembly of cell regulatory systems through protein interaction domains.** *Science* 2003, **300(5618):**445-52.
3. Zarrinpar A, Bhattacharyya RP, Lim WA: **The structure and function of proline recognition domains.** *Sci STKE* 2003, **2003(179):**.
4. Hanks SK, Quinn AM, Hunter T: **The protein kinase family: conserved features and deduced phylogeny of the catalytic domains.** *Science* 1988, **241(4861):**42-52.
5. Zeng G, Cai M: **Regulation of the actin cytoskeleton organization in yeast by a novel serine/threonine kinase Prk1p.** *J Cell Biol* 1999, **144(1):**71-82.
6. Pawson T: **Protein modules and signalling networks.** *Nature* 1995, **373(6515):**573-80.
7. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S, Quondam M, Zucconi A, Hogue CW, Fields S, Boone C, Cesareni G: **A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules.** *Science* 2002, **11(295(5553)):**321-4.
8. Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, Volkmer-Engert R, Cesareni G: **Protein interaction networks by proteome peptide scanning.** *PLoS Biol* 2004, **2(1):**E14.
9. Tonikian R, Zhang Y, Boone C, Sidhu SS: **Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries.** *Nat Protoc* 2007, **2(6):**1368-86.
10. Hutti JE, Jarrell ET, Chang JD, Abbott DW, Storz P, Toker A, Cantley LC, Turk BE: **A rapid method for determining protein kinase phosphorylation specificity.** *Nat Methods* 2004, **1(1):**27-9.
11. Beltrao P, Serrano L: **Comparative genomics and disorder prediction identify biologically relevant SH3 protein interactions.** *PLoS Comput Biol* 2005, **1(3):**e26.
12. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science* 2003, **302(5644):**449-453.
13. Henikoff JG, Henikoff S: **Using substitution probabilities to improve position-specific scoring matrices.** *Comput Appl Biosci* 1996, **12(2):**135-43.
14. Eddy SR: **Where did the BLOSUM62 alignment score matrix come from?** *Nat Biotechnol* 2004, **22(8):**1035-6.

15. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89(22):**10915-9.
16. McLachlan AD: **Repeating sequences and gene duplication in proteins.** *J Mol Biol* 1972, **64(2):**417-37.
17. Yaffe MB, Leparc GG, Lai J, Obata T, Volinia S, Cantley LC: **A motif-based profile scanning approach for genome-wide prediction of signaling pathways.** *Nat Biotechnol* 2001, **19(4):**348-53.
18. Obenauer JC, Cantley LC, Yaffe MB: **Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs.** *Nucleic Acids Res* 2003, **31(13):**3635-41.
19. Adamczak R, Porollo A, Meller J: **Accurate prediction of solvent accessibility using neural networks-based regression.** *Proteins* 2004, **56(4):**753-67.
20. Jones DT, Ward JJ: **Prediction of disordered regions in proteins from position specific score matrices.** *Proteins* 2003, **53(Suppl 6):**573-8.
21. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT: **The DISOPRED server for the prediction of protein disorder.** *Bioinformatics* 2004, **20(13):**2138-9.
22. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *J Mol Biol* 2001, **314(5):**1041-52.
23. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32(5):**1792-7.
24. Pei J, Grishin NV: **AL2CO: calculation of positional conservation in a protein sequence alignment.** *Bioinformatics* 2001, **17(8):**700-12.
25. Tonikian R, *et al.*: **Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins.** *PLoS Biol* 2009, **7(10):**e1000218.
26. Mok J, *et al.*: **Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation motifs.** *Sci Signal* 2010, **3(109):**ra12.
27. Turhan B, Bener A: **Analysis of Naive Bayes' assumptions on software fault data: An empirical study.** *Data Knowl Eng* 2009, **68(2):**278-290.
28. Puntervoll P, *et al.*: **ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins.** *Nucleic Acids Res* 2003, **31:**3625-3630.
29. Frank E, Hall M, Trigg L, Holmes G, Witten IH: **Data mining in bioinformatics using Weka.** *Bioinformatics* 2004, **20(15):**2479-81.

doi: 10.1186/1471-2105-11-243
Cite this article as: Lam *et al.*, MOTIPS: Automated Motif Analysis for Predicting Targets of Modular Protein Domains *BMC Bioinformatics* 2010, **11**:243