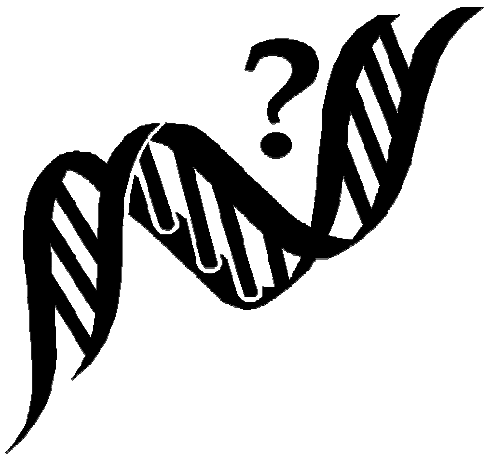


Automated Function Prediction, 2006

San Diego, CA

http://BioFunctionPrediction.org

Welcome!

On behalf of the Program Committee, the Scientific Committee and the organizers, we are happy to welcome you to San Diego for the second Automated Function Prediction conference, AFP 2006.

This year's program reflects the wide variety of exciting new approaches that have recently emerged for gene and protein function prediction. The Program Committee and the Scientific Committee were tasked with selecting 19 presentations from the 32 submitted. We have striven to provide a mix of scientific excellence and variety of approaches, and we believe that this effort is reflected in the program you are now holding.

Our keynote speakers will address the wide variety of problems associated with automated function prediction. New to this year, we have introduced a discussion panel engaging the audience with the keynote speakers. We are especially excited about discussing our plans for 2007, which include a Critical Assessment of Function Annotations, in the spirit of ongoing activities in the fields of computational structure prediction, and the prediction of protein-protein interactions. Please join us for this discussion on the second day of the conference.

Special thanks to John Wooley, Associate Vice Chancellor for Research at UCSD for securing conference funds and for his ongoing and enthusiastic backing of this endeavor.

Finally, we would like to thank everyone who supported this conference financially. AFP 2006 was made possible by grants from the NSF and the University of California Discovery grant, Calit2, our Gold Sponsors the Canadian Consulate in San Diego, the Canadian Consulate General in Los Angeles, and our Silver Sponsor TimeLogic.

We hope you will find AFP 2006 a stimulating and fruitful meeting.

Iddo Friedberg, Ph.D.
AFP 2006 co-chair

Adam Godzik, Ph.D.
AFP 2006 co-chair

AFP 2006 Program committee:

Christian Zmasek
Dana Weekes
Einat Sprinzak
Gill Bejerano
Jeffrey Chang
Lukasz Jaroszewski
Marco Punta
Michiel de Hoon
Mike Sanders
Naomi Cotton
Ora Furman
Rachel Kolodny
Sarah Boyd
Sourav Bandyopadhyay
S S Krishna
Thomas Hamelryck
Yanay Ofran
Yoav Freund
Yuzhen Ye

AFP 2006 Scientific Committee:

Adam Godzik
Ana Rodrigues
Barry Grant
Iddo Friedberg
Inbal Halperin
Martin Jambon

Conference Program booklet:

Dana Weeks
Iddo Friedberg

Conference Administrative Assistants:

Cindy Cook
Josephine Alaoen

Web site:

Iddo Friedberg
Ana Rodrigues
Naomi Cotton

Conference Program

Time	Speaker		Title	Page
8:30-8:45			Opening remarks	
Day 1 - Wednesday, August 30, 2006				
9:00-9:45	Adam Godzik	Burnham Institute for Medical Research and the University of California San Diego	TBA	
9:45-10:10	Ying Xu		High-Resolution Functional Assignments of Genes through Mapping KEGG Pathways to Bacterial Genomes	
10:10-10:35	Hon Nian Chua		Guilt by Indirect Functional Association	
10:35-11:00	Break			
11:00-11:45	Christos Ouzounis	European Bioinformatics Institute	CorrIE: probabilistic proteins sequence annotation based on functional classifications	
11:45-12:10	Jason Li		Gene function prediction based on synteny and discriminative learning	
12:10-12:30	Emmanuel Perrodou		Proteome annotation with MACSIMS (Multiple Alignment of Complete Sequences Information Management System)	
12:30-2:00	Lunch			
2:00-2:45	Steven E. Brenner	University of California, Berkeley	Problems and Proposals in Protein Molecular Function Prediction	
2:45-3:10	Barbara E. Engelhardt		SIFTER: a Model of Molecular Function Evolution to Predict Protein Function	
3:10-3:35	Shawn Cokus		An Improved Method for Identifying Functionally Linked Proteins Using Phylogenetic Profiles	
3::35-4:00	Break			
4:00-4:25	Kimmen Sjölander		Structural Phylogenomic Analysis: New Methods and Challenges	
4:25-4:50	Edgar Vallejo		Inferring Functional Coupling of Genes from Phylogenetic Profiles using the Bond Energy Algorithm	
4:50-5:35	Russ B. Altman	Stanford University	Clustering Protein Microenvironments for	

			Automated Function Prediction	
5:30-8:00	poster session & dinner			
Day 2 - Thursday, August 31, 2006				
08:30-09:15	Anna Tramontano	University of Rome, La Sapienza	Revisiting Function Prediction at CASP	
09:15-09:40	Guy Nimrod		Identification of DNA-binding Proteins Based on the Projection of Evolutionary Conservation on their Structures: Progress Report	
09:45-10:10	Thomas Funkhouser		Protein Function Prediction by Matching Volumetric Models of Active Sites	
10:10-10:35	Sean D. Mooney		Supervised Classification of Enzyme Residue Function using Machine Learning Methods	
10:35-11:00	Break			
11:00-11:45	Shoshana Wodak	University of Toronto	Identifying meaningful functional modules in the yeast protein protein interaction network	
11:45-12:10	Andreas Henschel		Function and Interaction Prediction Using Multiple Motif Descriptors for Classified Domain-Domain interactions and ligand binding sites.	
12:10-12:30	T. M. Murali		Hierarchically Consistent Prediction of Gene Function	
12:30-2:00	Lunch			
2:00-2:45	Terry Gaasterland	Scripps Institute of Oceanography	Function Prediction in the RNA World	
2:45-3:10	Jian Qiu		Kernels for Protein Structures	
3:10-3:35	Eugene Ie		Multi-Class Protein Classification using Adaptive Codes	
3:35-4:00	Break			
4:00-5:30	Discussion Panel	Keynote Speakers		
06:00 PM	Banquet			
Day 3 - Friday, September 01, 2006				
08:30-9:15	Philip Bourne,	University of California, San Diego	Novel Ways to Think about Protein Structure and its Impact on Function Prediction	
9:15-9:40	Lei Xie		A Robust and Efficient Algorithm for Geometrical Characterizations of Protein Structures and Its applications in Studying	

			Protein-Ligand Interactions	
9:40-10:05	Marc A. Marti-Renom		The AnnoLite Program for Rapid and Reliable Comparative Annotation of Protein Structures	
10:05-10:35	Break			
10:35-11:00	Tsai-Tien Tseng		Voltage-gated Ion Channels and Auxiliary Subunits in the Transport Classification System	
11:00-11:25	*Bill Chang, Glyn Roberts, Jo-Ming Ong, Saman Halgamuge and Nalin Wickramarachchi*		Protein Motif Discovery with Particle Swarm Algorithm	
11:30	Concluding remarks			

Talk Abstracts

KEYNOTE

TBA

Adam Godzik,
Burnham Institute for Medical Research and the University of California, San Diego

High-Resolution Functional Assignments of Genes through Mapping KEGG Pathways to Bacterial Genomes

Fenglou Mao, Hongwei Wu, and Ying Xu*
Department of Biochemistry and Molecular Biology and Institute of Bioinformatics
University of Georgia, Athens, GA 30622
*To whom correspondence should be addressed: xyn@bmb.uga.edu

1. INTRODUCTION

We have developed a computational capability for mapping KEGG metabolic pathways to sequenced bacterial genomes. This capability assigns genes of a bacterial genome to specific enzymatic roles of a given KEGG pathway using a two-level strategy: (a) initial assignment is based on the premise that a bacterial metabolic pathway is in general encoded by a number of (in general transcriptionally co-regulated) operons and based on predicted functions of individual genes possibly at a low-resolution level; and (b) filling the gaps, the unassigned enzymes, in a partially-assigned pathway based on the detected co-evolutionary, co-occurrence and co-regulated relationships and predicted protein-protein interactions between un-annotated genes and genes already assigned to the pathway. To facilitate automated functional assignment of genes, we have developed a number of supporting computational tools, including prediction of operons [1], uberoperons [2] and regulons (unpublished results).

A. Initial KEGG pathway mapping: We have developed a computational algorithm for mapping a KEGG pathway to a specified bacterial genome. The algorithm starts by searching each gene in the target genome against gene databases with annotated functions such as the nr database and making functional predictions, possibly at a low-resolution level, based on identified homology relationship. Then a number of genes (possibly zero) with annotated functions will be predicted as possible candidates for each enzyme in the KEGG pathway, based on the match between the predicted gene functions and the enzyme. We then assign at most one candidate gene to each enzyme of the KEGG pathway, using the following criteria: the overall consistency between the predicted gene functions and their assigned enzymatic roles should be as high as possible, and the selected genes should be clustered as much as possible as the predicted operons. This problem is formulated as a constrained optimization problem, specifically a linear integer programming problem, and solved using a commercial linear integer programming solver COIN. This overall prediction capability has been implemented as a computer program, called PMAPKEGG. Using this capability, we have mapped over 140 KEGG pathways to 300+ sequenced bacterial genomes, including E. coli, for which detailed validation has been done using pieces of information from multiple sources. For every sequenced bacterial genome, our mapping results cover a substantial fraction of all the genes in that genome. Detailed data will be reported in an extended version of this abstract.

It should be noted that the operon prediction for each target bacterial genome is made using three prediction programs, JPOP [1], OFS [3] and VIMSS [4]. A simple majority-vote scheme is used for the final operon prediction. In the actual formulation of the problem, we have also taken into consideration the predicted uber-operon information using our own prediction program [2], where a uber-operon represents a group of opreons whose union is conserved across multiple genomes, which gives a higher prediction coverage than using operons alone.

B. Filling gaps in a partially assigned pathway: The mapped KEGG pathways often contain "gaps", unassigned enzymatic roles, due to various reasons. We have developed a computational procedure attempting to fill in these gaps, using three types of information: (a) co-evolutionary and co-occurrence information between assigned genes and unassigned & un-annotated genes, (b) predicted regulon information (i.e., transcriptionally co-regulated operons), and (c) protein protein interaction information derived using various techniques such as the gene fusion method. It has been generally known that co-evolutionary, co-occurrence and co-regulation information of genes can help to predict functional relatedness among genes, even when functions of some of the genes are unknown. By employing this idea, we have recently developed a computational technique for predicting genes that are possibly working

closely together in the same biological process [5,6]. Using this capability, we have predicted an initial set of candidate genes for each "gap" in a partially assigned KEGG pathway. We have then predicted protein-protein interaction relationships between the candidate genes for each "gap" with the genes already assigned to the network neighborhoods of the gap. Our final prediction for each gap is selected, using a trained neural network, based on the predicted functional relatedness and protein-protein interaction. We found that we were able to make correct gene assignments (as top assignments), for about 30% of the gaps, on a large test set using well characterized E coli. pathways after manually removing some of the assigned genes (1-3 genes are randomly removed from each assigned pathway). Detailed results will be reported in an extended version of this abstract.

Concluding remarks: By assigning genes of a bacterial genome to KEGG pathways, we can provide functional prediction of genes at a high-resolution level (knowing exactly the functional role in a well understood metabolic pathway), compared to the low-resolution functional annotation typically provided by a genome annotation system, e.g., gene A encodes a protease, and also can assign un-annotated genes to possible functional roles in a metabolic pathway. Our computational prediction program consists of a number of prediction and analysis tools, which are pipelined together to facilitate large-scale applications. A database containing all mapped KEGG pathways to each of the 300+ sequenced bacterial genomes is currently being developed, and will be made publicly available within a few months. This collection of mapped pathways has provided a very rich set of information for studies of bacterial metabolic pathways and their evolution. For example, by comparing the same mapped KEGG pathways across multiple genomes, we can derive information about how a pathway has evolved in adaptation to an organism's living environments, leading to general information about pathway evolution and adaptation.

Acknowledgement: This work was supported in part by the by National Science Foundation (NSF/DBI-0354771, NSF/ITR-IIS-0407204) and the US Department of Energy's Genomes to Life Program.

2. REFERENCES

1. X. Chen, Z. Su, P. Dam, B. Palenik, Ying Xu and T. Jiang, Operon prediction by comparative genomics: an application to the Synechococcus WH8102 genome, Nucleic Acids Research, 32 (7), 2147 – 2157, 2004.
2. D. Che, G Li, F. Mao, H Wu, and Ying Xu, "Detecting uber-operons in microbial genomes", Nucleic Acids Research, 2006 (in press).
3. BP. Westover, JD. Buhler, JL Sonnenburg, and J.I. Gordon, Operon prediction without a training set, Bioinformatics, 21, 880-888, 2005.
4. MN. Price, KH. Huang, EJ. Alm and AP. Arkin, A.P. A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res, 33, 880-892, 2005.
5. F. Mao, Z. Su, V. Olman, P. Dam, Z. Liu and Ying Xu, Mapping of Orthologous Genes in the Context of Biological Pathways: an Application of Integer Programming, Proc Natl Acad Sci USA, 103, 129-134, 2006.
6. H. Wu, Z. Su, V Olamn, Ying Xu, Prediction of functional modules through comparative genome analysis and application of gene ontology, Nucleic Acids Research, 33, 2822-2837, 2005.

Guilt by Indirect Functional Association

Hon Nian Chua, Wing-Kin Sung, Limsoon Wong*
National University of Singapore, School of Computing, 3 Science Drive 2, Singapore 117543
*To whom correspondence should be addressed: wongls@comp.nus.edu.sg

1. INTRODUCTION

Although sequence similarity search has been proven useful in many cases, it has fundamental limitations. First, only a fraction of newly discovered sequences have identifiable homologous genes in the current databases. Second, the most prominent vertebrate organisms in GenBank have only a fraction of their genomes present in finished sequences. New bioinformatics methods allow inference of protein function using ``associative analysis'' of functional properties to complement the traditional sequence homology based methods. Associative properties that have been used to infer function not evident from sequence homology include: co-occurrence of proteins in operons or genome context; proteins sharing common domains in fusion proteins; proteins in the same pathway; proteins with correlated gene expression patterns; etc.

Most approaches (1,2,5,6) in predicting protein function from protein-protein interaction data utilize the observation that a protein often share functions with proteins that interact with it (its level-1 neighbors). However, proteins that interact with the same proteins (i.e. level-2 neighbors) may also have a greater likelihood of sharing similar physical or biochemical characteristics with the target protein. We are interested to find out how significant is functional association between level-2 neighbors and how they can be exploited for protein function prediction. We also investigate how to integrate protein interaction information with other types of information to improve the sensitivity and specificity of protein function prediction, especially in the absence of sequence homology (results omitted due to lack of space). We have been investigating graph-based methods to this problem (3). We report some newly obtained results here, especially for fruit fly and roundworm.

We observe that there are proteins that do not share any function with their immediate interaction partners (i.e., level-1 neighbors, S1) and yet share some function similarity with the interaction partners of their immediate partners (i.e., level-2 neighbors, S2). In fact, when we briefly inspect the yeast protein-protein interaction data downloaded from the GRID database (4), we find that out of the 4162 annotated proteins, only 1999 or 48.0% share some function with its level-1 neighbours. Of the remaining proteins, 943 share some similarity with at least one of its level-2 neighbours, making up around 22.7% of the ORFs. Less than 2% of the annotated proteins share functions exclusively with level-1 neighbours. We have repeated our experiments on two other species, *Drosophila melanogaster* (fruit fly) and *Caenorhabditis elegans* (roundworm). Although interaction data and annotations for these two species are less comprehensive, the results are consistent with that for *Saccharomyces cerevisiae* (yeast); see Table 1. Assuming that there is no unobserved interaction or annotation, *indirect functional association* would be a reasonable explanation for this observation. Such an indirect functional association can be considered as an instance of the "guilt by association" principle --- the common "property" between the level-2 neighbors and the target protein that is used for deriving the "association" is precisely the set of proteins that they both interact with, namely the level-1 neighbors; it is plausible that two proteins that interact with a common set of proteins have a good likelihood of sharing similar physical or biochemical characteristics, and thus exhibit a common function.

Genome	Annotation	S_1-S_2	S_2-S_1	$S_1 \cap S_2$	$S_1 \cup S_2$
S. cerevisiae	MIPS	0.007193	0.226574	0.463960	0.706872
D. melanogaster	GO	0.008801	0.168622	0.138138	0.315561
C. elegans	GO	0.007193	0.051237	0.061080	0.119510

Table 1. Fraction of annotated yeast proteins that share function with 1) level-1 neighbours exclusively; 2) level-2 neighbours exclusively; 3) level-1 and level-2 neighbours; and 3) level-1 or level-2 neighbours for different species and annotation schemes.

Some approaches (6) have suggested using the common interacting partners between two proteins as an estimate of their functional similarity. We devise a measure, *Functional Similarity Weight* (FS-Weight):

$$S_{FS}(u,v) = \frac{2|N_u \cap N_v|}{|N_u - N_v| + 2|N_u \cap N_v| + \lambda_{u,v}} \times \frac{2|N_u \cap N_v|}{|N_v - N_u| + 2|N_u \cap N_v| + \lambda_{v,u}}$$

where *Np* refers to the set that contains *p* and its level-1 neighbours, and $\lambda u,v$ is included in the computation to penalize similarity weights between protein pairs when any of the proteins has very few level-1 neighbors.

Using the FS-Weight measure, we propose a weighted averaging method, *FS Weighted Averaging*, to predict the function of a protein based on the functions of the level-1 and level-2 neighbors. The likelihood that a protein *p* has a function *x* is estimated by:

$$f_x(u) = \frac{1}{Z}\left[\lambda\pi_x + \sum_{v \in N_u}\left(S_{TR}(u,v)\delta(v,x) + \sum_{w \in N_v} S_{TR}(u,w)\delta(w,x)\right)\right]$$

where *STR(u,v)* is the Transitive FS-Weight score for *u* and *v; δ(p, x) = 1* if *p* has function *x*, *0* otherwise; πx is the frequency of function *x* in annotated proteins; $0 \leq \lambda \leq 1$ is the weight representing the contribution of background frequency to the score; and Z is the sum of all weights.

To study whether prediction performance of FS-Weighted Averaging is superior to existing methods, leave-one-out cross validation is performed using 3 prediction techniques (a) FS-Weighted Averaging; (b) Chi-Square of level-1 neighbors (5); and (c) Neighbour Counting of level-1 neighbors (6), on the *S. cerevisiae*, *D. melanogaster* and *C. elegans* genomes. The precision versus recall graphs are presented in Figure 1. The FS-Weighted Averaging method consistently outperforms the other two in all cases.

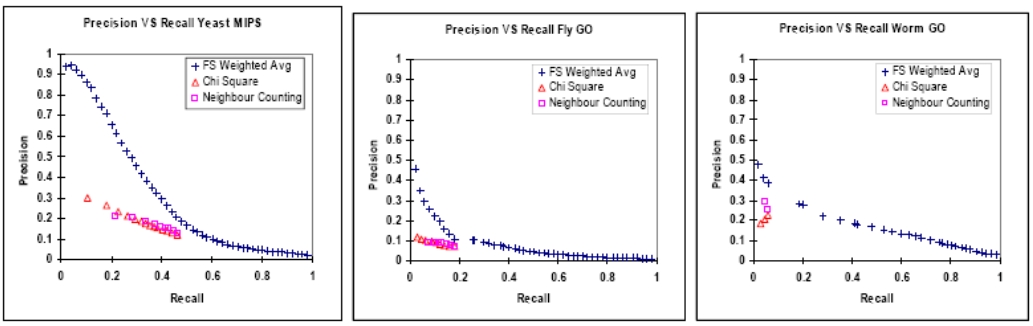


Figure 1. Precision vs Recall graphs for function prediction using (a) FS Weighted Averaging; (b) Chi-Square; and (c) Neighbour Counting for *S. cerevisiae* (yeast), *D. melanogaster* (fly) and *C. elegans* (worm) with functional annotations from MIPS and GO.

3. REFERENCES

1. Brun et al. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol*. 5(1):R6, 2003.
2. Deng et al. Prediction of protein function using protein-protein interaction data. *JCB*, 10(6):947-960, 2003.
3. Chua et al. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, accepted.
4. Breitkreutz et al. The GRID: the General Repository for Interaction Datasets. *Genome Biol*, 4(3):R23, 2003.
5. Hishigaki et al. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, 18(6):525-531, 2001.
6. Schikowski et al. A network of interacting proteins in yeast. *Nature Biotechnology*, 18:1257-1261, 2000.

KEYNOTE

CORRIE: Probabilistic Protein Sequence Annotation Based on Functional Classifications

Levy ED [1,2], Gilks WR [3], Audit B 4, Goldovsky L [1,5], Ouzounis CA* [1,5,6]
1 Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK • 2 Computational Genomics Group, MRC Laboratory of Molecular Biology, Hills Rd, Cambridge CB2 2QH, UK • 3 Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR, UK • 4 Laboratoire Joliot-Curie and Laboratoire de Physique, CNRS UMR5672, Ecole Normale Supérieure, 46 Allée d'Italie, F-69364 Lyon CEDEX 07, France • 5 Computational Genomics Unit and • 6 Institute of Agrobiotechnology, Center for Research & Technology Hellas, PO Box 361, GR-57001 Thessalonica, Greece – (2, 5, 6: current addresses)
*To whom correspondence should be addressed: ouzounis@certh.gr

1. INTRODUCTION

The explosion of sequencing technologies has resulted in an ever-increasing gap between the discovery of new gene sequences and their experimental characterization. The accumulation of raw sequence data has dictated the use of computational techniques for the inference of their possible functional roles, based on evolutionary conservation of structure and function. However, this widely used, empirical process has not been considered as a fundamental problem in computational biology that requires rigorous analysis.

The typical solution to annotation transfer involves the inference of functional properties based on sequence similarity (1). This procedure can be divided into two steps: (i) the establishment of a list of proteins of known function and significant sequence similarity to the uncharacterized sequence (2); (ii) the selection of those characterized sequences from which the annotation might be transferred (3). This type of annotation process relies on the assumption of a strong relationship between protein sequence and function. This hypothesis is generally fair (4), even though many studies have demonstrated the existence of counter examples that can lead to annotation errors (5, 6, 7, 8). Two major classes of errors can be distinguished: (i) the short-listed homologous protein(s) have a different function from the sequence to be annotated (erroneous assignment, despite correct reference); (ii) the transferred annotations were themselves not correct (erroneous reference, despite correct assignment). The second class of errors along with the iterative usage of annotation transfer gives rise to the specific problem of error propagation when newly annotated sequences are included in the reference database used for the homology search. Simulation studies have shown that dramatic consequences on the reliability of database annotations are likely to arise from this process, with detrimental consequences for the quality and integrity of reference databases (9). In order to improve our control over these two classes of errors, it would be very useful to associate a measure of reliability to the annotations obtained (3). In this way, we might limit the introduction of new errors and their propagation by not admitting the transfer of the less reliable annotations, according to specific criteria.

We address this issue by developing a probabilistic framework to the homology-based annotation process. Our approach relies on the usage of a reference dataset, where protein sequences are pre-classified into (an arbitrary number of) functional classes (10). Here, an assignment corresponds to a membership in a specific functional class; thus, function sharing becomes an explicit property. The possibility for a protein to perform a particular function is then assessed based on its similarity relationships with all protein sequences known to perform this function. This enables, for instance, the unambiguous consideration of both presence and absence of similarity. Note that most existing methods map functions to proteins, first by "clustering" proteins based on sequence similarities (irrespectively of any function sharing), and second by combining available functional descriptions in the (most relevant) cluster to annotate the uncharacterized sequence(s) (11, 12, 13). An innovative feature of our strategy is that a given sequence is mapped to a functional class of characterized proteins, reversing the logic of mapping a functional class to a group of similar sequences. We introduce Correspondence Indicators (CIs) as a novel measure to quantify the relationship between a protein sequence and a functional class. A CI results from the combination of pairwise similarity scores between a query sequence of interest and all the members of a functional class. In this implementation, we use the BLAST bit-scores as a measure of pairwise similarity but other measures can be used in principle.

We subsequently define univariate (one functional class at a time) and multivariate (all functional classes simultaneously) Bayesian schemes for sequence annotation based on CIs. Importantly, this methodology provides probabilistic estimates for query sequences to belong to a functional class. This approach allows the possibility to filter out dubious functional predictions on the basis of appropriate, user-defined criteria. To test this new strategy, we have previously re-annotated a database of 28,088 enzyme proteins (catalyzing one reaction only and belonging to Enzyme Classification (EC) categories with more than 11 members). We re-mapped each enzyme to its 4-digit EC category in a fully automated manner, using the two Bayesian schemes, and the BLAST best-hit method as a reference. Both the univariate and multivariate Bayesian schemes outperform the traditional approach: at the highest confidence level, they exhibit small and similar error rates, $r \sim 0.0021$ and $r \sim 0.0020$ respectively, for a high coverage of the database (90.6% and 96.0% respectively). These rates compare favorably to those obtained with the BLAST best-hit annotation method, where the error rate is, respectively, 1.5X and 2X higher, for identical coverage. To achieve error rates equal to the above with the BLAST best-hit method, the threshold for admissible annotations needs to be modified for higher specificity, while coverage drops to 54% and 51% respectively. We have now implemented this strategy into a software program, called CORRIE which stands for Correspondence Indicator Estimation, and we announce its availability for wider use by the computational biology community. The software takes as input a reference set of protein sequences, their association to a (functional) classification and an all-vs-all similarity table. Then, for any unclassified sequence, CORRIE generates a probability for its membership to each of the functional classes. The software will be made available for downloading in 2006 through the following URL: http://www.genomes.org/cgg/Services.html. A key feature of our methodology is the quantification of the reliability of annotations; the assignment probability represents an attractive candidate, both versatile and compact, to qualify non-experimentally based, inferred annotations (10). In principle, it could be taken into account by the Bayesian annotation framework allowing its iterative usage, without risking the propagation of annotation errors (9). It is our hope that the Bayesian annotation strategy will contribute to more robust automatic annotation pipelines.

2. REFERENCES

1. M. A. Andrade, C. Sander. Bioinformatics: from genome data to biological knowledge. Curr. Opin. Biotechnol., 8:675-683,1997.
2. C. A. Ouzounis, P. D. Karp. The past, present and future of genome-wide re-annotation. GenomeBiology 3, c2001.1-c2001.6,2002.
3. P. D. Karp. What we do not know about sequence analysis and sequence databases. Bioinformatics, 14:753-754,1998.
4. C. A. Wilson, J. Kreychman, M. Gerstein. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J. Mol. Biol., 297:233-249,2000.
5. N. C. Kyrpides, C. A. Ouzounis. Whole-genome sequence annotation: 'Going wrong with confidence'. Mol. Microbiol., 32:886-887,1999.
6. P. Bork, E. V. Koonin. Predicting functions from protein sequences--where are the bottlenecks? Nature Genet., 18:313-318,1998.
7. D. Devos, A. Valencia. Intrinsic errors in genome annotation. Trends Genet., 17:429-431,2001.
8. J. A. Gerlt, P. C. Babbitt. Can sequence determine function? Genome Biol., 1:REVIEWS0005,2000.
9. W. R. Gilks, B. Audit, D. De Angelis, S. Tsoka, C. A. Ouzounis. Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics, 18:1641-1649,2002.
10. E. Levy, C. A. Ouzounis, W. R. Gilks, B. Audit. Probabilistic annotation of protein sequences based on functional classifications. BMC Bioinformatics 6, 302,2005.
11. F. Abascal, A. Valencia. Automatic annotation of protein function based on family identification. Proteins, 53:683-692,2003.
12. W. G. Krebs, P. E. Bourne. Statistically rigorous automated protein annotation. Bioinformatics, 20:1066-1073,2004.
13. A. M. Leontovich, L. I. Brodsky, V. A. Drachev, V. K. Nikolaev. Adaptive algorithm of automated annotation. Bioinformatics, 18:838-844,2002.

Gene Function Prediction Based on Synteny and Discriminative Learning

Jason Li*, S. K. Halgamuge, Sen-Lin Tang
Bioinformatics Section, Dynamic Systems & Control Group, DOMME,
University of Melbourne, VIC 3010, Australia
*To whom correspondence should be addressed: lij@mame.mu.oz.au

1. INTRODUCTION

We present a system capable of predicting gene functions based on gene context information. The system contains two automated learning units with distinct roles: a novel synteny-based clustering technique to identify closely related genomes and a Support Vector Machine (SVM) for discriminative classification on gene functions.

The background of our work lies in bacteriophage genomics. Bacteriophages are known to be the most abundant living entities on earth with an immense amount of novel proteins yet to be discovered (1). Finding protein homologs among phages using sequence similarity based methods is often insufficient in identifying functionally similar genes as they can exhibit no sequence similarity. On the other hand, the conservation of gene order, or synteny, has played a significant role in helping to infer gene functions within bacteriophage genomes (2). Conservation of synteny is a concept used in comparative genomics to analyse closely related species. It has helped to identify genes and regulatory elements in a variety of studies including yeasts (3) and mammals (4). Existing software can visualise synteny relationships across multiple species to assist biologists, but they lack an automated process to identify functions of uncharacterised genes (5)-(6). To automate the synteny analysis process for function predictions, we have developed a system called SynFPS, or the **Syn**teny-based **F**unction **P**rediction **S**ystem. Although our experimental focus is on bacteriophages, the system can be applied to other areas where synteny arguments are suitable and large collection of genomes are available.

An overview of SynFPS is shown in Figure 1. The genome annotation database as shown in the figure defines the scope of analysis for the system. In our current work, it consists of 296 phage genomes, which were retrieved automatically from GenBank via the NCBI common gateway interface using SynFPS. The system begins by identifying in the database a collection of genes that correspond to a set of user-specified gene functions. Each gene function is specified in the form of a regular expression pattern (7), which may contain multiple keywords that represent the same function. The use of regular expression is aimed at tackling annotation discrepancies among coding sequences in databases that do not have vocabulary control. SynFPS provides visual aids for manual inspection over the genes identified by the system. In this step, users may include additional genes or exclude mistaken genes.
SynFPS employs a synteny-scoring system to rate the relatedness among genomes. The scoring system is conceptually similar to a high-level pairwise alignment, where genes in one genome are aligned to their corresponding genes in another. Each score is associated with five factors: the degree of conservation of gene order, locations of the genes relative to the length of the genome, absolute locations of the genes in terms of basepairs, and relative and absolute distances between consecutive aligned genes. If the differences in these factors between two genomes are large, then the alignment score will be low. These differences are rated by a Gaussian function for sensitivity control. The system then employs an adaptive k-mean clustering algorithm to cluster the genomes into different groups based on their synteny resemblances reflected in the pairwise scores. The clusters of genomes are analysed separately and individually in the last stage of the system. For each cluster, we use the information of the previously identified genes to predict the functions of other genes that exhibit similar context. This is achieved by extracting a set of genes from the cluster and converting them into positive and negative training data for a discriminative classification. Positive data are formed by the group of previously identified genes, with each gene function representing one class. Negative data comprise the genes that are neighbours to the positive genes. The size of neighbourhood is determined by the statistics of the gene locations in that particular cluster. We use 99% confidence interval on the gene locations of each class to determine the range in which neighbour genes are to be included. This interval also determines the set of candidate genes on which function predictions are performed. SynFPS uses a Support Vector Machine (SVM) (8) for the classification. For each gene

function, the SVM produces a binary result on each candidate gene indicating whether or not the gene belongs to that function class. Since the number of gene functions is specified by the user and is not likely to cover every possible function, only a subset of the candidate genes – those with positive results – will eventually be assigned with predicted functions. Information and downloads related to SynFPS can be accessed via http://www.synteny.net.

2. FIGURES

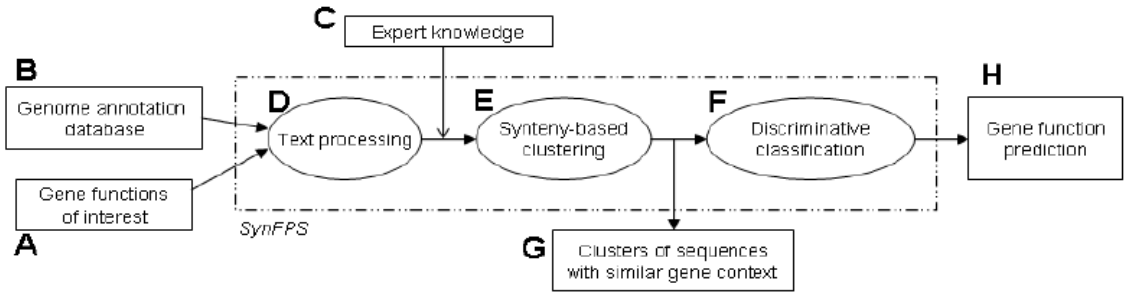


Figure 1. Structure of the Synteny-based Function Prediction System (SynFPS). The dotted line represents the system boundary, outside which lies the system inputs and outputs. A set of gene functions (A) specified in the form of regular expressions are matched against the genome database (B) via the text processing unit (D), which result may then be refined (C). A clustering system (E) based on the synteny scores of the matching genes brings together genomes that are evolutionarily or syteny-related (G). Such information is used to generate a set of positive and negative data (genes) to train the classification system (F) that produces function prediction results (H).

3. REFERENCES

1. Kutter, E. and Sulakvelidze, A. 2005. *Bacteriophages: Biology and Applications* : CRC Press.
2. Brüssow, H. and Hendrix, R.W. 2002. Phage genomics: Small is beautiful. *Cell* 108:13-16.
3. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423(6937):241-254.
4. Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I., Hardison, R.C. 2003. Cross-species sequence comparisons: a review of methods and available resources. *Genome Research* 13:1-2.
5. Pan, X., Stein, L., Brendel, V. 2005. SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics* 21(17):3461-3468.
6. Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., Dubchak, I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Research* 32:273-279.
7. Sipser, M. 2005. Chapter 1: Regular Languages. *Introduction to the Theory of Computation* : PWS Publishing. pp. 31-90.
8. Keerthi, S.S., Shevade, S.K, Bhattacharyya, C. and Murthy, K.R.K. 2001. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computing* 13:637-649.

Proteome Annotation with MACSIMS (Multiple Alignment of Complete Sequences Information Management System)

Julie D. Thompson[1], Odile Lecompte[1], Arnaud Muller[2], Emmanuel Perrodou[1], AndrewWaterhouse[3], Jim Procter[3], Geoffrey J. Barton[3], Frédéric Plewniak[1], Patrice Koehl[4], and Olivier Poch[1]

[1] Laboratoire de Biologie et Genomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, (CNRS/INSERM/ULP), BP 163, 67404 Illkirch Cedex, France.

[2] The Laboratory of Molecular Biology, Genetic Analysis & Modelling, 84, rue du Val Fleuri L-1526, Luxembourg.

[3] Post Genomics & Molecular Interactions Centre, School of Life Sciences, University of Dundee, Scotland.

[4] University of California, Davis, One Shields Avenue, Davis CA 95616, USA

1. INTRODUCTION

The application of high-throughput techniques such as genomics, proteomics or transcriptomics means that vast amounts of heterogeneous data are now available in the public databases. New integrated management systems are now needed for data collection, validation and analysis. Hierarchized multiple alignments of complete sequences (MACS) with defined protein family and subfamily level organisation provide an ideal environment for the integration of this mass of information. By placing the sequence in the framework of the overall family, MACS can be used to identify important functional, structural or cellular motifs or domains that have been conserved through evolution but also can highlight particular non-conserved features resulting from specific events or perturbations.

We have developed MACSIMS (http://bips.u-strasbg.fr/MACSIMS/), a MACS-based information management system that combines knowledge-based methods with complementary ab initio sequence-based predictions for protein family annotation and analysis. MACSIMS takes advantage of the recently developed Multiple Alignment Ontology (MAO) (Thompson et al. Nucleic Acids Res 2005, 33:4164-4171), to integrate different types of data in the framework of the multiple alignment. A wide range of information, from taxonomic data and functional descriptions to individual sequence features, such as structural domains, secondary structures and active site residues, is mined from public databases. In addition, different prediction programs are run for each sequence to identify low complexity segments, hydrophobicity properties, potential transmembrane helices, coiled coil segments, etc. New algorithms have been developed for reliable data validation, consensus predictions and rational propagation of information from the known to the unknown sequences. The reliability of the MACSIMS data management system has been demonstrated using a largescale test set of 83 automatic alignments based on a structural benchmark database. The 83 alignments contained a total of 10250 PDB or Uniprot sequences with a total of 150772 sequence features (functional definitions, Interpro, GO, taxon entries, secondary structure elements …) retrieved from the databases and a further 2535 predicted features. The final number of sequence features (261791) was increased by 70% through the propagation of the validated features to unknown sequences. For each of the alignments, the sequence features propagated by MACSIMS, such as functions, domains or active site motifs, were compared to the known sequence features. The specificity in these tests was shown to be above 99%, and the sensitivity was 91%. MACSIMS has been integrated in various high-throughput projects ranging from structural genomics (SPINE project) for target selection and characterization up to analysis of proteins involved in human genetic diseases (MS2PH project). In addition, an original integrated annotation of the Mycobacterium smegmatis genome has been performed taking advantage of the MACSIMS XML format output files (http://www-bio3d-igbmc.u-strasbg.fr/macsim.dtd) that provide a structured format for automatic annotation, validation and data comparison at a proteome level. These applications illustrate the data integration potential of MACSIMS and demonstrate that application of new developments in ontology-based methods facilitates intelligent knowledge extraction and decision support for structural, functional or evolutionary annotation and analyses.

KEYNOTE

Problems and Proposals in Protein Molecular Function Prediction

Steven E. Brenner
University of California, Berkeley, CA 94720-3102
brenner@compbio.berkeley.edu

1. INTRODUCTION

We are adrift in a sea of protein function predictions. The sequences of over 10^7 proteins are known, to which have been attributed millions of functions. However, the GO Annotation database lists only about 50,000 proteins whose annotation had—at a minimum—a human examine the prediction. The number of proteins recorded as experimentally characterized is even smaller. Yet, biologists depend upon protein function annotations for insight and analysis.

How accurate are automated function predictions? Few automated annotators are so bold as to make an estimate of accuracy. Indeed, due to assessment more by quantity rather than quality, it appears that the number of false positive function predictions has increased as the genome era has progressed. A decade-old study of annotations in *M. genitalium*, found that the minimum error rate between three groups was at least 8% and estimated actual error rates were ~2-3% higher (1). Automated predictions of function for the adenine/adenosine/AMP deaminases were recently assessed; the test evaluated programs' ability to accurately describe proteins whose experimentally characterization is reported in the literature, but not present in electronic databases. Common and established function prediction programs made errors on about $1/3^{rd}$ of the proteins (2).

A decade ago, when the protein sequence databases were small and largely manually curated, Eugene Koonin estimated that most errors in protein function are actually propagations of existing database errors. That is, the function of the protein to be annotated is the same as that of the matched database protein, but the protein in the database had been incorrectly described. It was widely appreciated that this problem could be largely overcome by having every protein annotation supported by traceable evidence. This would allow each new protein annotation to be associated with a degree of confidence, and would allow propagation of corrections to follow propagated errors. The GO Annotation database now incorporates the GO evidence codes, and it provides information for millions of proteins (3). The task of incorporating all literature evidence into the databases is immense and ongoing. Function prediction methods that incorporate evidence would seem less prone to error propagation.

One of the most subtle problems in protein function prediction is functional change between very closely related homologs. In these cases, it is hard to imagine any function prediction method providing the precisely correct protein function. Expert manual annotators often have sufficient expertise to recognize such cases, and to provide a more generic functional prediction, for example "amino acid permease," when the top BLAST hit is a "histidine permease" (1). Recent automated methods have incorporated similar ideas, by making predictions at intermediate nodes of GO rather than at the leaves (e.g., 4). These approaches are promising, though it remains to be seen whether the GO DAG is a satisfactory proxy for evolutionarily accessible functional variability.

Several researchers have proposed genosperology or phylogenomics as a powerful approach for meeting many the other challenges of protein function prediction (5). These methods use a full reconciled phylogenetic history of a protein family to make protein function predictions, rather than a subset of high-scoring BLAST hits. This approach has many advantages. First, the most similar sequences according to BLAST may not be those with the most recent common ancestor, and thus are not those which are most likely to share a common function. This problem grows worse as more data are added to a tree, meaning that the BLAST highest-hits approach is systematically flawed and may yield increasingly erroneous results as data increase. Use of a phylogeny incorporates the evolutionary history explicitly and directly. Second, a phylogeny suggests an evolutionarily-principled means of integrating functional evidence, even when data

are sparse or noisy. While originally applied manually, phylogenetically-motivated protein function prediction has now been deployed in automated methods. Orthostrapper uses bootstrapping to identify reliable orthologs, and uses only those for protein function prediction (6). As such orthostrapper makes relatively few annotations when restricted to experimental evidence, but those predictions it makes are rarely erroneous. SIFTER (Statistical Inference of Function Through Evolutionary Relationships) uses a statistical model of functional evolution and makes predictions supported by a posterior probability, allowing it to incorporate information throughout an evolutionary tree to make predictions for every protein even when data are limited and problematic (7). Interestingly, in addition to using provided protein functions as evidence, SIFTER can also evolutionarily integrate predictions from other programs. In its authors' hands, SIFTER makes more correct predictions of protein function on the deaminases than any other method, and independent tests are needed.

2. REFERENCES

1. Brenner SE. 1999. Errors in genome annotation. *Trends in Genetics 15*:132-133.
2. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. 2005. Protein molecular function prediction by Bayesian phylogenomics. *PLoS Computational Biology 1*:e45. doi:10.1371/journal.pcbi.0010045.
3. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, et al. 2004. The Gene Ontology annotation (GOA) database: Sharing knowledge in UNIPROT with Gene Ontology. *Nucleic Acids Res 32*:262-266.
4. Eisen JA 1998. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res 8*:163-167.
5. Martin DMA, Berriman M, Barton GJ. 2004. GOtcha: A new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics 5*:178-195.
6. Storm CE, Sonnhammer EL. 2002. Automated ortholog inference from phylogenetic trees and calculation of ortholog reliability. *Bioinformatics 18*:92-99.
7. Engelhardt BE, Jordan MI, Brenner SE. 2006. A statistical graphical model for predicting protein molecular function. *Proceedings of the 23rd International Conference on Machine Learning. in press.*

SIFTER: A Model of Molecular Function Evolution to Predict Protein Function

Barbara E Engelhardt*, Michael I Jordan, Steven E Brenner
Computer Science Division, University of California, Berkeley
EECS Department, University of California Berkeley, Berkeley, CA, 94720, USA
Department o f Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA
*To whom correspondence should be addressed: bee@cs.berkeley.edu

1. INTRODUCTION

SIFTER automatically predicts protein molecular function using a statistical model that describes molecular function evolution in a particular protein family. The methodology has its roots in phylogenomics as described by Eisen (1998), and it has similar robustness properties and ability to generalize across a wide range of protein families. We automate phylogenomics by applying statistical methods to propagate functional information within a graphical model that uses a reconciled gene phylogeny as its structure.

SIFTER's statistical model encodes general knowledge of how molecular function evolves within a sequence-based phylogenetic tree. Inputs are a query protein sequence and any available function characterizations for proteins homologous to the query protein. The result of statistical inference is a set of posterior probabilities for each protein; these are used to predict the molecular function of that protein. Our results focus on the application of SIFTER to three different protein families.

Ongoing work in several areas based on the SIFTER methodology includes application to a full fungal genome, protein functional characterization experiments, and an information theoretic method to select the optimal protein to experimentally characterize from the set of unannotated proteins in a family.

As with basic manual phylogenomics, many decisions go into the data processing and integration for this model of function prediction. These decisions include our choice of Gene Ontology (GO) (Ashburner et al., 2002) to represent molecular function terms, and in particular attempting to differentiate between a set of functional terms at the more specific levels of the ontological hierarchy. We select a set of homologous proteins using Pfam (Bateman et al., 2002) domains because of their common association with molecular functions. We use parsimony to reconstruct the phylogeny and Forester to reconcile the phylogeny with the species tree (Zmasek & Eddy, 2001), and our method appears fairly robust to the given reconciled phylogeny. Finally, we use the GO Annotation database (GOA) (Camon et al., 2002) to derive a list of possible functions from the set of all GO molecular function terms, and use this as input to the graphical model.

We present results from applying our model to two protein families, and compare our prediction results on the extant proteins to other available protein function prediction methods: BLAST (Altschul et al., 1990), GOtcha (Martin et al., 2004), and Orthostrapper (Storm & Sonnhammer, 2002).

For a gold standard dataset from the AMP/adenosine deaminase family, our method achieves 93.9% in prediction accuracy where related methods BLAST achieves 72.7%, GOtcha achieves 87.9%, and Orthostrapper achieves 72.7% in prediction accuracy. The AMP/adenosine deaminase family is particularly interesting because a subset of the proteins in the family have an additional protein domain conferring growth factor activity as a second molecular function. The method of evaluation we prefer for this family is a Receiver Operating Characteristic (ROC)-type analysis. This type of analysis measures the rate of false positives against the rate of false negatives and enables a comparison of the methods' ability to rank both functions highly for multifunction proteins, not just the ability to rank one of the two functions highly. From this analysis, SIFTER predicts function better than the three other methods across almost all false positives rates.

A second family that is particularly interesting is the aminotransferase family, because a significant amount of homoplasy, or convergent evolution, has occurred within this family. We use a gold standard dataset based on the GOA database, a manual literature search, and a series of experimental characterizations performed by our collaborator. Sparse database annotations prevent BLAST, GOtcha, and Orthostrapper from correctly predicting that any protein in the family has tyrosine substrate specificity (as opposed to aspartate), all achieving 66.7% accuracy (8 of 12). When SIFTER is applied to this family with the gold standard dataset as input, it achieves 75% prediction accuracy (9 of 12), getting only one of the four TATase predictions incorrect. On a larger, unpublished gold standard dataset, SIFTER performs equally well on the standard input and also using the ratio of AATase to TATase activity for each characterized protein as input in lieu of the preferred activity alone.

2. REFERENCES

1. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. J Mol Biol 215: 403 410. Curr Opin Struct Biol 10: 366 370.
2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2002) Gene Ontology: Tool for the uni cation of biology. The Gene Ontology Consortium. Nat Genet 25: 25 29.
3. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, et al. (2002) The Pfam protein families database.Nucleic Acids Res 30: 276 280. 63.
4. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, et al. (2004) The Gene Ontology annotation (GOA) database: Sharing knowledge in UNIPROT with Gene Ontology. Nucleic Acids Res 32: 262 266.
5. Eisen J.A. (1998) Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8: 163-167.
6. Engelhardt, B.E., Jordan, M.I. and Brenner, S.E. 2005. Protein molecular function prediction through Bayesian phylogenomics. PLoS Computational Biology 1:e45.
7. Engelhardt, B.E., Jordan, M.I. and Brenner, S.E. 2006. A graphical model for predicting protein molecular function. Proceedings of the 20th Annual International Conference on Machine Learning (ICML), to appear.
8. Martin D.M.A., Berriman M, Barton G.J. (2004) GOtcha: A new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 5: 178-195.
9. Storm C.E., Sonnhammer E.L. (2002) Automated ortholog inference from phylogenetic trees and calculation of ortholog reliability. Bioinformatics 18: 92-99.
10. Zmasek C.M., Eddy S.R. (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 17: 821 828.

An Improved Method for Identifying Functionally Linked Proteins Using Phylogenetic Profiles

Sayaka Mizutani, Shawn Cokus, Matteo Pellegrini*
Department of Molecular, Cell and Developmental Biology
University of California, Los Angeles
*To whom correspondence should be addressed: matteop@mcdb.ucla.edu

1. INTRODUCTION

We propose a novel approach to improve Phylogenetic Profile analysis. The method considers phylogenetic relationships between genomes to improve estimates of profile similarities.

To date about 300 bacterial genomes have been fully sequenced. Although these sequences provide us a wealth of information, the functions of the products of many of the genes have yet to be characterized. Development of methodologies that can predict their function is an important goal for bioinformatics. The most widely used methods for protein function prediction are based on the detection of homologies based on sequence alignments. However, these approaches are often insufficient, as many proteins have no functionally characterized homologs. Moreover, it is not possible to completely define the function of an isolated protein, for function depends intimately on contextual information, such as the interactions, pathway and cellular localization a protein is associated with.

Functional characterization of proteins using phylogenetic profiles has emerged as an important technique during the past few years1. A phylogenetic profile is a data structure that is assigned to each protein within a genome and whose elements indicate the presence and the absence of homologs of the protein in another genome (see Figure 1). The underlying assumption of methods that utilize these profiles is that proteins that function together tend to co-occur across organisms. Thus, clusters of proteins with similar profiles correspond to pathways and complexes, and participation in such a cluster may be used as evidence that an uncharacterized protein shares this function.

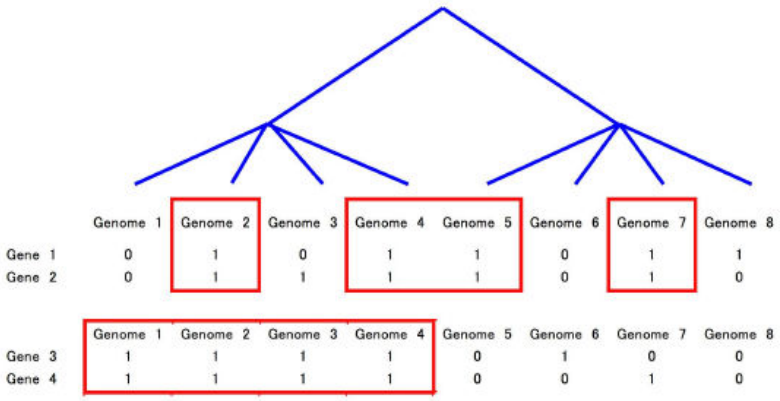
Various metrics have been used to quantify the similarity between two phylogenetic profiles including the Hamming distance1, probability of matches using the hypergeometric distribution2 and mutual information3. However, these metrics do not consider the underlying phylogeny of the genomes in the profile. As Figure 1 suggests, there is ample reason to believe that accounting for phylogeny should improve our ability to detect truly co-evolving genes (Genes 1 and 2), from those that are merely present in a subset of related genomes (Genes 3 and 4). Although some methods have been developed to account for genome phylogeny when scoring profile similarities4,5, the complexity of their implementation has limited the application of these approaches. Here we present a simple heuristic that partially accounts for the relationships between organisms when scoring profile pairs.

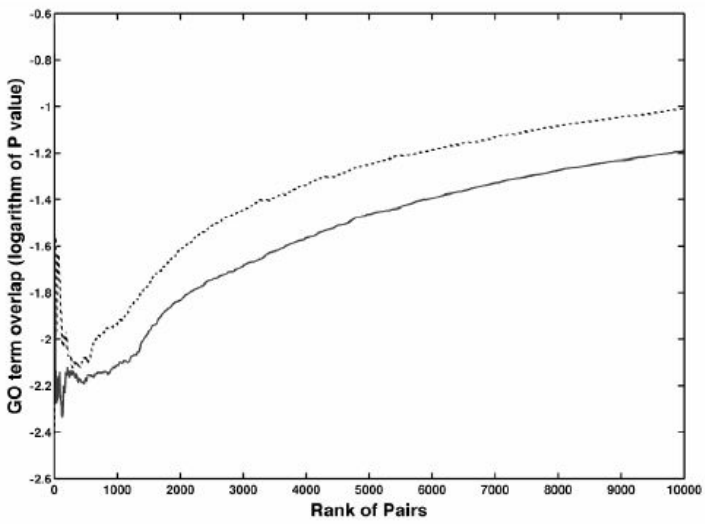
Our approach involves two components. The first is to account for runs of consecutive matched homologs in phylogenetic profiles to distinguish between conservation of occurrences within clusters of related organisms versus conservation across disparate species. The second involves an extension of the hypergeometric distribution that accounts for the number of proteins in each genome. Each of these components is described by a simple analytical formula and they are easily combined to yield a single probability that two profiles are significantly similar. We test this approach by measuring how often proteins in significant profile pairs share the same Gene Ontology terms. As shown in Figure 2, by ranking protein pairs based on the significance of their profile similarity, we see that the average p-value for two proteins sharing a GO term is more significant when we account for runs and genome size (solid line) than if we don't (dashed line).

In conclusion, we have developed a simple heuristic method that accounts for genome phylogenies when computing phylogenetic profile similarities. We have shown that this approach improves our ability to reconstruct various pathways including those involved in ribosomal biogenesis and RNA degradation. In

the future we plan to incorporate this new methodology into the Prolinks database (http://mysql5.mbi.ucla.edu/cgi-bin/functionator/pronav).

2. FIGURES





3. REFERENCES

1. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96, 4285-8 (1999).
2. Wu, J., Kasif, S. & DeLisi, C. Identification of functional links between genes using phylogenetic profiles. Bioinformatics 19, 1524-30 (2003).
3. Date, S. V. & Marcotte, E. M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat Biotechnol 21, 1055-62 (2003).
4. Vert, J. P. A tree kernel to analyse phylogenetic profiles. Bioinformatics 18 Suppl 1, S276-84 (2002).
5. Barker, D. & Pagel, M. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput Biol 1, e3 (2005).

Structural Phylogenomic Analysis: New Methods and Challenges

Kimmen Sjölander,* Dan Kirshner, Nandini Krishnamurthy and Duncan Brown
Berkeley Phylogenomics Group, University of California, Berkeley
*To whom correspondence should be addressed: kimmen@berkeley.edu

1. INTRODUCTION

Phylogenomic inference of protein function, combining phylogenetic tree construction, integration of experimental data, and differentiation of orthologs and paralogs, has been shown to reduce the annotation errors associated with function prediction by homology, and to improve the accuracy of functional classification. The explicit integration of structure prediction and analysis in this framework, which we call *structural phylogenomics*, provides additional insights into protein superfamily evolution, and improves function prediction accuracy (1,2).

Because phylogenomic inference, even in the absence of structural considerations, is quite complicated, the Berkeley Phylogenomics Group has developed a number of software tools and web servers to enable the use of these methods by biologists. As of April 22, 2006, the Berkeley Phylogenomics Group Universal Proteome Explorer (3) has almost 10,000 "books" for protein families and domains, and over 700K hidden Markov models (HMMs) enabling classification of user-submitted sequences to families and subfamilies. These phylogenomic HMM libraries enable classification of proteins to functional subfamilies (4), prediction of protein structure, active site and subfamily-specificity residues, and cellular localization. To date, almost 2,000 unique users across the world have visited our resource, with thousands of hits monthly. All the data in the resource is downloadable, including multiple sequence alignments, phylogenetic trees, subfamily and family HMMs, and predicted subfamilies. The pipeline developed to construct the Universal Proteome Explorer is shown in Figure 1.

Our analysis of database annotation errors suggests that a large fraction of existing errors can be traced to annotation transfer based on local (i.e., non-global) homology. To avoid these errors, we have developed the FlowerPower global homolog clustering server (5). In our Universal Proteome Explorer phylogenomic resource, this clustering of proteins into domain architecture equivalence classes is followed by phylogenetic analysis. These orthogonal analyses are designed to ensure that annotation transfer take place under carefully controlled conditions.

Identification of functional subfamilies facilitates the organization and annotation of a protein family. When experimental data is available for one or more members of a family, supervised learning approaches (e.g., SVM methods) can be used to automatically classify novel sequences to functional subtypes. In many cases, however, experimental data is not available and unsupervised learning approaches based on analysis of protein structure or amino acid sequence are required in order to predict the inherent subtypes in a family. We employ several approaches for this task, including identification of conserved phylogenetic clades and subfamily identification using the SCI-PHY algorithm (6). SCI-PHY uses agglomerative clustering and Dirichlet mixture densities to construct a phylogenetic tree, and employs minimum description length principles to cut the tree into subtrees. Our analyses show SCI-PHY subfamilies correspond closely to functional subtypes found by biologist experts and also to conserved clades found through standard phylogenetic analyses (see Figure 2).

In this talk, I will describe the pipeline used to construct the Universal Proteome Phylogenomic Explorer, present a few of the core methods developed by my lab for phylogenomic inference, and discuss some of the successes we have had using these approaches to detect and correct existing annotation errors. Finally, I will describe some of the challenges we see to large-scale automated assignment of protein function using phylogenomic approaches.

Acknowledgements: This work is funded by an RO1 from the NHGRI and by a Presidential Early Career Award for Scientists and Engineers from the National Science Foundation.

2. FIGURES

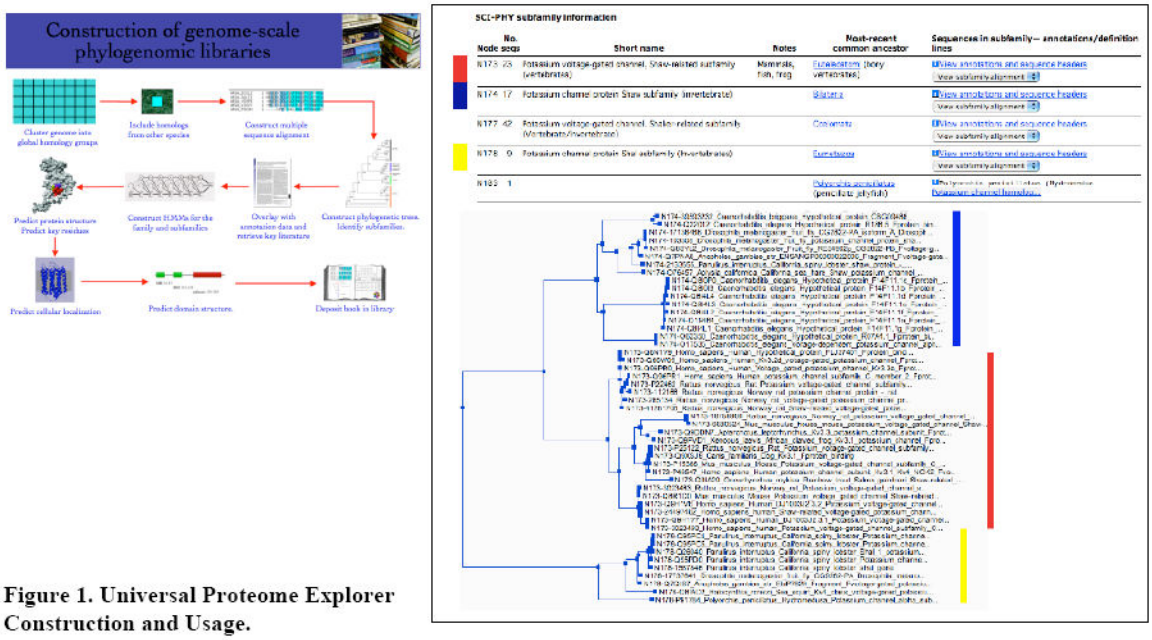


Figure 1. Universal Proteome Explorer Construction and Usage.

Left: Pipeline for phylogenomic library construction. Globally alignable homologs are clustered using the FlowerPower algorithm. MSAs are constructed using MUSCLE and masked prior to phylogenetic tree construction. GO annotations and evidence codes are retrieved and overlaid on the tree topology. Subfamilies are identified using the SCI-PHY algorithm, and HMMs are constructed for the family as a whole and also for each subfamily. Other bioinformatics analyses are performed, and the protein family "book" is deposited in the library for biologists to browse and for classification of user-submitted sequences.

Right: Functional subtype identification using SCI-PHY subfamily identification and phylogenetic tree analysis. Shown here is the Universal Proteome Explorer book Voltage-gated K+ Shaker/Shaw family (bpg000014). *Top*: SCI-PHY subfamilies, based on a minimum description length decomposition of the family into subfamilies under a Dirichlet mixture density. *Bottom*: Close-up of a Maximum Likelihood tree constructed for this family using the PhyML software. The Universal Proteome Explorer includes both standard trees (using Neighbor-Joining, Maximum Likelihood and Parsimony methods) and SCI-PHY subfamilies for user inspection. Shaded bars added manually to show the correlation between SCI-PHY subfamilies and conserved phylogenetic clades.

3. REFERENCES

1. Sjölander, K. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*. 2004 Jan 22;20(2):170-9.
2. Brown D and Sjölander, K, "Functional classification using phylogenomic inference." To appear in *PLoS Computational Biology*.
3. http://phylogenomics.berkeley.edu/UniversalProteome/
4. Brown D, Krishnamurthy N, Dale J, Christopher W, and Sjölander, K "Subfamily HMMs in Functional Genomics", *Proceedings of the Pacific Symposium on Biocomputing,* 2005.
5. http://phylogenomics.berkeley.edu/flowerpower/
6. Sjölander, K, "Phylogenetic inference in protein superfamilies: Analysis of SH2 domains," *Proceedings of the Conference Intelligent Systems for Molecular Biology* 1998 6:165-74.

Inferring Functional Coupling of Genes from Phylogenetic Profiles using the Bond Energy Algorithm

Ryosuke Watanabe[1], Edgar Vallejo[1]*, Enrique Morett[2]
[1]Computer Science Department, ITESM Campus Estado de México
Carr. Lago de Guadalupe km 3.5, Atizapán de Zaragoza, Edo. Mex., 529226, México
[2] Department of Cellular Engineering and Biocatalysis, IBT, UNAM
Av. Universidad 2001, Cuernavaca, Morelos, 62210, México
*To whom correspondence should be addressed: vallejo@itesm.mx

1. INTRODUCTION

The development of automated methods for determining functional coupling of genes from sequence and genomic data has becoming an increasingly important area of investigation in genomics and computational biology. In effect, the determination of unknown gene interactions in functional pathways and perhaps, their association with disease relies crucially on sound computational algorithms capable of producing meaningful predictions.

Among modern postgenomic approaches developed in recent years, those based on the correlated presence and absence of genes among a collection of organisms have proven to be particularly effective (4). Theoretically, with the increasing availability of complete genomes from more organisms, these methods hold the promise of increasing efficiency. Particularly, phylogenetic profiles have been used for assigning protein function, for localizing proteins in cells, and for reconstructing metabolic pathways, among other applications (2,8).

The efficacy of predictions obtained from phylogenetic profiles depend critically on the employed clustering method. Most clustering algorithms used to date are based on the determination of Euclidean and Hamming distances, which means that the clustering is directed by the intrinsic properties of these patterns and no additional information is often considered, although there are a few exceptions (5,7). Further, these algorithms often require *a priori* information on the number of clusters and the initial positions of centers.

The Bond Energy Algorithm (BEA) has been used for years in diverse areas such as distributed database design and production research (1). BEA constructs an $N \times N$ affinity matrix A in which entry ajk indicate how closely related is element j with element k. The algorithm then rearranges the rows/columns of A so that they are closely related as possible to their neighboring rows/columns. This produces an arranged matrix whose underlying structure could be interpreted by inspection. Moreover, clustering could be easily performed on this matrix by selecting an arbitrary number of entries in the diagonal such that the affinity measure among the grouped elements is maximized.

Most importantly, in grouping a pair of rows/columns together, BEA considers not only their intrinsic similarity, but also the strength of the relationship with respect to a third-party element. This aspect of the algorithm is very important in applications such as automated function prediction, in which multiple relationships (with varying degree) among elements are expected.

We conducted experiments using BEA for clustering phylogenetic profiles obtained from the Cluster of Orthologous Groups of proteins, COGs database (6). For the preliminary experiments shown here, we used six phylogenetic profiles from the class K (transcription), five from the class J (translation, ribosomal structure and biogenesis), and five from the class M (cell envelope biogenesis, outer membrane). We compared the results produced by BEA and the *k*-means clustering algorithm (3). Figure 1 and 2 shows the results by BEA and *k*-means, respectively.

Experimental results indicate that BEA is capable of producing accurate grouping of functional related proteins from their phylogenetic profiles. In contrast, the *k*-means algorithm failed to produce meaningful predictions. Further, the algorithm produces a conformation of clusters that is visually compelling to the

naked eye. Appropriate representations of protein relationships are of increasing importance as more orders of magnitude (hundreds and thousands of elements) are increasingly typical in automated function prediction experiments.

2. FIGURES

	COG1594K	COG1761K	COG2012K	COG1644K	COG1996K	COG1243K	COG2870M	COG0859M	COG1043M	COG0815M	COG0787M	COG1190J	COG0223J	COG1185J	COG0242J	COG0193J
COG1594K	26	26	26	26	26	24	11	9	7	4	3	4	1	1	0	1
COG1761K	26	26	26	26	26	24	11	9	7	4	3	4	1	1	0	1
COG2012K	26	26	26	26	26	24	11	9	7	4	3	4	1	1	0	1
COG1644K	26	26	26	26	26	24	11	9	7	4	3	4	1	1	0	1
COG1996K	26	26	26	26	26	24	11	9	7	4	3	4	1	1	0	1
COG1243K	24	24	24	24	24	26	11	9	7	6	3	6	3	3	2	3
COG2870M	11	11	11	11	11	11	26	24	22	19	18	15	14	16	15	14
COG0859M	9	9	9	9	9	9	24	26	24	19	20	17	16	18	17	16
COG1043M	7	7	7	7	7	7	22	24	26	21	20	17	18	20	19	18
COG0815M	4	4	4	4	4	6	19	19	21	26	21	18	21	23	22	21
COG0787M	3	3	3	3	3	3	18	20	20	21	26	19	22	24	23	22
COG1190J	4	4	4	4	4	6	15	17	17	18	19	26	23	21	22	23
COG0223J	1	1	1	1	1	3	14	16	18	21	22	23	26	24	25	26
COG1185J	1	1	1	1	1	3	16	18	20	23	24	21	24	26	25	24
COG0242J	0	0	0	0	0	2	15	17	19	22	23	22	25	25	26	25
COG0193J	1	1	1	1	1	3	14	16	18	21	22	23	26	24	25	26

Figure 1. Clusters produced by the bond energy algorithm

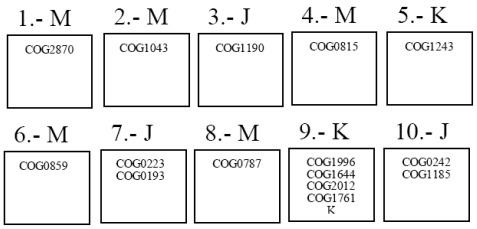


Figure 2. Clusters produced by the *k*-means algorithm

3. REFERENCES

1. Arabie, P. and Hubert, L. The bond energy algorithm revisted. *IEEE Transactions on systems, man, and cybernetics*, 20:1, pp.268-274.

2. Chen, L., Vitkup, D. Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biology 2006.*

3. Duda, R.O., Hart, P.E and Store, D.C. *Pattern Classification. Second Edition*. Wiley-Interscience, 2000.

4. Pellegini, M., et al Asigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *PNAS 1999*

5. Sun, J. et al. Refined phylogenetic profiles method for predicting protein-protein interactions. *Bioinformatics Vol. 21 No. 16 2005 pp. 3409-3415.*

6. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science*. 1997 Oct 24;278(5338):631-7.

7. Yamanishi, Y., et. al. Extraction of groups from phylogenetic profiles using independent component analysis. *Genome Informatics 2002 vol. 13 PP 61-70*

8. Wu, J., et. al. Identification of functional links between genes using phylogenetic profiles. *Bioformatics Vol. 19 No. 12 2003 pp. 1524 – 1530.*

KEYNOTE

Clustering Protein Microenvironments for Automated Function Prediction

Sungroh Yoon[1], Jessica Ebert[2], Russ B. Altman[2]*
[1]Computer Systems Laboratory, Stanford University, Stanford, CA 94305, USA
[2]Department of Genetics, Stanford University, Stanford, CA 94305, USA
*To whom correspondence should be addressed: russ.altman@stanford.edu

1. INTRODUCTION

For high-throughput function prediction on protein structures, we have developed a computational method to cluster residues in a non-redundant subset of the Protein Data Bank (PDB) (1) based on their physicochemical environments as calculated by FEATURE, a robust statistical method for representing models of sites in macromolecules (2). The hope is to be able to discover residues whose environments are similar enough that we can infer a of relationship among their functional/structural roles.

Figure 1 outlines the proposed approach. From approximately 9,600 non-redundant structures in the PDB (50% similarity threshold; available at http://www.rcsb.org/pdb/clusterStatistics.do), we derived 3.6 million FEATURE vectors, each of which is a 264-dimensional vector of (mostly discrete and a few continuous) variables. Each dimension represents a certain protein microenvironment such as those shown in Figure 4(c). As a preprocessing step, each FEATURE vector was converted into its binary form. The entire set of binary vectors was then clustered recursively by 2-means clustering until the size of every cluster became less than a given threshold (we used 10,000). We found about 600 such clusters, and each cluster was further clustered into smaller sub-clusters by hierarchical clustering. Approximately 15,000 clusters were found after this two-step clustering procedure.

The distance metric used is called *F-distance*, which is defined in Figure 2(a). This distance metric is based upon the notion of the information coefficient. For dimension i, its information coefficient f_i implies how far that dimension is from randomness. Figure 2(b) shows the distribution of the information coefficients over the entire 3.6 million FEATURE vectors used.

Figure 3 compares three different distance metrics used for clustering 15 previously known FEATURE clusters in terms of the silhouette value (3). A silhouette value can quantify clustering quality of each object in a cluster by a continuous number between +1 (perfectly clustered) and -1 (the opposite). As seen in Figure 3, using the F-distance gives a higher median silhouette value or a best clustering result than using the Euclidean or Hamming distance.

From the clusters discovered by the two-step clustering procedure, we selected only those that contain FEATURE vectors with identical SCOP (4) and PROSITE (5) annotations and, if relevant, Enzyme Commission numbers. The residues involved in these clusters may then prove to be involved in inferring functional specificity. The annotation of a cluster corresponds to the predicted function of the FEATURE environment characterized by the cluster. Figure 4 shows an example of the FEATURE environment discovered by our approach. Figure 4(a) represents the insulin environment in terms of binary values, and Figure 4(b) shows those dimensions that are significantly different from noise with respect to their information coefficients.

We implemented the proposed clustering approach as an on-line algorithm, and its running time and memory usage were very reasonable in most cases.

2. REFERENCES

1. Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., et al. 2002. The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* 58: 899-907.
2. Altman, R.B., Liang, M. and Wu, S. 2005. The FEATURE System for Protein Structure Annotation, AFP 2005.
3. Kaufman, L. and Rousseeuw, P.J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*: Wiley, New York.

4. Murzin, A.G., Brenner, S.E., Hubbard T. and Chothia C. 1995. SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J. Mol. Biol.* 247: 536-540.
5. Hulo N., Bairoch A., Bulliard V., Cerutti L., De Castro E.,Langendijk-Genevaux P.S., Pagni M. and Sigrist C.J.A. 2006. The PROSITE database. *Nucleic Acids Res.* 34:D227-D230.

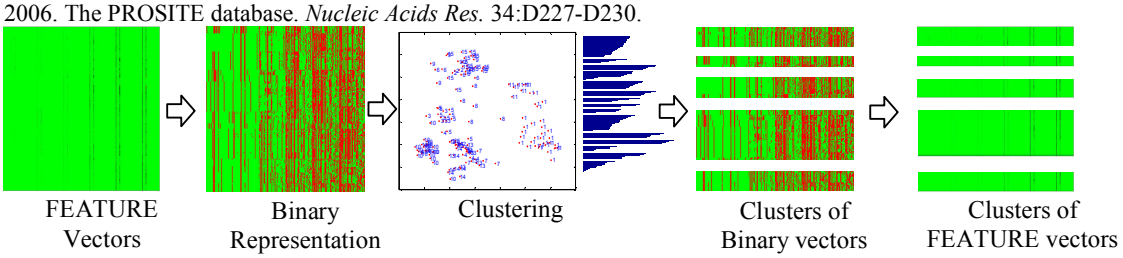
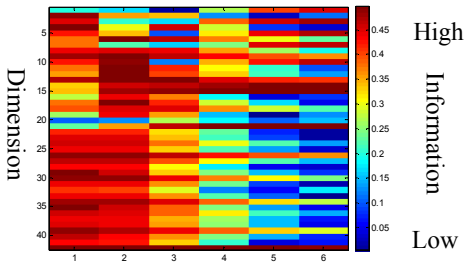


Figure 1: Overview of the proposed method.

$$f_i = \left| 0.5 - \frac{\sum_n X_i}{n} \right|$$

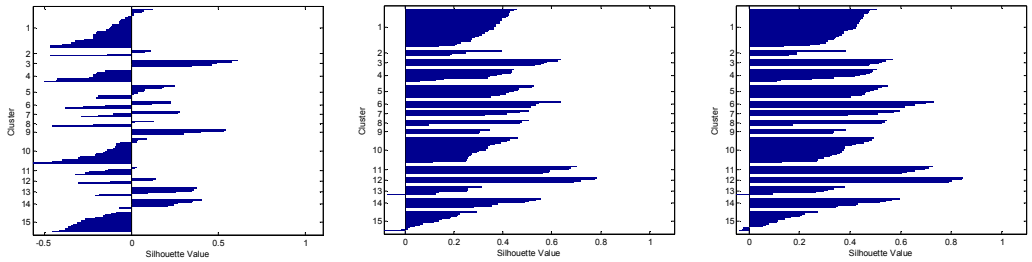
$$F\text{-}distance = \frac{1}{n} \sum_i 2 f_i |X_i - Y_i|$$

(a) Information coefficient f_i and F-distance (n is the total number of FEATURE vectors; X_i and Y_i represent the i^{th} coordinate of FEATURE vectors X and Y in their binary form, respectively.)



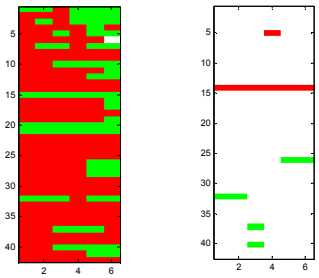
(b) Distribution of f_i

Figure 2: Defining distance metric for clustering.



(a) Euclidean (S = -0.12) (b) Hamming (S = 0.35) (c) F-distance (S = 0.41)

Figure 3: Comparison of different metrics used for clustering 15 known clusters of FEATURE vectors in terms of the median silhouette value (*S*).



(a) FEATURE environment for insulin

(b) Significant absence (red) and presence (green)

1 Aliphatic carbon
2 aromatic carbon
3 carbon w/partial +charge
4 aliphatic C next a polar atom
5 amide carbon
6 carboxyl carbon
7 amide nitrogen
8 +charged nitrogen
9 aromatic nitrogen
10 amide oxygen
11 carboxyl oxygen
12 hydroxyl oxygen
13 sulfur
14 partial charge
15 Van der Waals volume
16 charge
17 negative charge
18 positive charge
19 charge on histidines
20 hydrophobicity
21 solvent accessibility
22 residue name is ala
23 residue name is arg
24 residue name is asn
25 residue name is asp
26 residue name is cys
27 residue name is gln
28 residue name is glu
29 residue name is gly
30 residue name is his
31 residue name is ile
32 residue name is leu
33 residue name is lys
34 residue name is met
35 residue name is phe
36 residue name is pro
37 residue name is ser
38 residue name is thr
39 residue name is trp
40 residue name is tyr
41 residue name is val
42 residue name is other

(c) The description of each dimension used

Figure 4: An example of the FEATURE environments discovered.

KEYNOTE

Revisiting Function Prediction at CASP

Paolo Marcatili, Marialuisa Pellegrini-Calace, Simonetta Soro and Anna Tramontano*
Department of Biochemical Sciences, University "La Sapienza", P.le Aldo Moro, 5, 00185 Rome, Italy
*To whom correspondence should be addressed: anna.tramontano@uniroma1.it

1. INTRODUCTION

The ability to predict the function of a protein, given its sequence and/or three-dimensional structure, is an essential requirement for exploiting the wealth of data made available by the genomics and structural genomics projects and is therefore raising increasing interest in the computational biology community. In order to foster developments in the area as well as to establish the state of the art of present methods, a function prediction category was tentatively introduced in the 6th edition of the CASP (Critical Assessment of Techniques for Protein Structure Prediction) worldwide experiment (1). The assessment of the performance of the methods was made difficult by at least two factors: on one side the experimentally determined function of the targets was not available at the time of assessment (2), on the other the experiment is ran blindly and this did not permit to verify whether the convergence of different predictions towards the same functional annotation was due to the similarity of the methods or to a genuine signal detectable by different methodologies. Recently, we collected information about the methods used by the various predictors and revisited the results of the experiment by verifying how often and in which cases a convergent prediction was obtained by methods based on different rationale and propose a method for classifying the type and redundancy of the methods (3). Our results demonstrate that predictions deriving from consensus of different methods can reach an accuracy as high as 80% (3). It follows that some of the predictions submitted to CASP6, once re-analysed taking into account the type of converging methods, can provide very useful information to experimentalists interested in the function of the target proteins.

One conclusion that can be derived from our analysis is that a key to success is the use of a combination of methods based on independent criteria.

We ourselves are investigating the possibility of using protein – protein interaction maps (4-9) to this end. The first step, however, must be the assessment of the reliability of the information contained in interaction maps. At a molecular level protein protein interactions can be mediated by domain interactions, by domain binding to short recognition motifs or by low-complexity sequences. We are developing a method that, through a systematic analysis of sequences, structures and/or models of potentially interacting protein pairs, identifies and assigns a quantitative confidence rate to such interactions. The rationale consists in identify potential interactors of the same protein (hub), look for sequence and/or structural similarities among them and associate a p-value to each interaction.

Preliminary results show that even sequence similarity alone is able to validate a sizeable fraction of the interaction pairs, ranging from 15% to 40% in different datasets. We are now in the process of integrating this first step with structure-based methods.

2. REFERENCES

1. Moult, J., Fidelis, K., Rost B., Hubbard T and Tramontano, A. 2005. Critical Assessment of Methods of Protein Structure Prediction (CASP) - Round VI. Proteins. 7 pp. 3-7.
2. Soro, S. and Tramontano, A. 2005 Function Prediction at CASP6. Proteins. vol. 7 pp. 214-224 3. Soro, S., Pellegrini-Calacae, M, and Tramontano,.A. 2006 Revisiting the prediction of protein function at CASP6. FEBS Jin press
4. Uetz, P. et al. 2000 A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Nature, 403, pp. 623–627
5. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. 2001 A comprehensive two hybrid analysis to explore the yeast protein interactome. Proc. Natl Acad. Sci. USA, 98, pp 4569–4574

6. Gavin, A.C. et al. 2002 Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 415,pp 141–147
7. Krogan, NJ, et al. 2006 Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440:pp 637-43
8. Butland, G, Peregrin-Alvarez, J.M , Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J. and Emili, A. 2005 A Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature 433 pp. 531–537.
9. Ho, Y. et al. 2002 Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature, 415, pp. 180–183

Identification of DNA-Binding Proteins Based on the Projection of Evolutionary Conservation on their Structures: Progress Report

Guy Nimrod, Nir Ben-Tal*
Department of Biochemistry, George S. Wise Faculty of Life Science, Tel Aviv University, Ramat Aviv 69978, Israel
*To whom correspondence should be addressed: NirB@tauex.tau.ac.il

1. INTRODUCTION

DNA-binding proteins take part in a variety of crucial cellular functions, i.e., DNA translation, maintenance and replication. Although many DNA-binding proteins have been biochemically identified, there is a large number of proteins, especially from the structural-genomics projects, that might have DNA-binding properties, but were not thus identified. Detection of such proteins is therefore a major goal of the structural bioinformatics discipline. Computational methods to this effect used the characteristic electrostatic potentials of the DNA-binding site (e.g. 1-3), specific structural motifs (3), evolutionary conservation (4-5) and geometric attributes of the protein surfaces (1-2, 6).

We recently presented PatchFinder, an algorithm for the identification of contiguous functional regions (patches) in proteins of known 3-dimensional (3D) structure (7). PatchFinder was based in the observation that functionally-important amino acid positions are often found to be evolutionarily conserved amongst sequence homologues and are exposed to the solvent (8). This is in contrast with structurally important positions, which tend to be conserved but buried within the protein core. It is also noticed that these residues are found in close proximity to each other on the protein surface (e.g. 9-11). The PatchFinder methodology is based on the following steps: (a) The assignment of an evolutionary conservation score (12) to each amino-acid position in the protein; (b) In order to approximately discriminate between positions that are conserved due to structural constraints and those that are conserved due to their functional importance, the patch search procedure is limited to positions that are exposed to the solvent; (c) Extraction of the maximum-likelihood (ML) patch of conserved residues. This patch is considered to be the main functional site of the protein.

Since PatchFinder was published, we improved it by extraction of contiguous residues using the Delaunay triangulation (13,14) as implemented in the qhull software package (15) rather than the simple criterion of distance between atoms that was originally used (7). In the current work, we examine the ability of the improved PatchFinder to predict DNA-binding sites in proteins. PatchFinder was tested on a dataset of 60 representative protein/double-stranded DNA (dsDNA) complex structures (1). Each entry in the dataset is related to a unique SCOP (16) family or has a sequence identity cutoff of <40% to each of the other members in the dataset. In addition, each entry has at least 4 homologous sequences in the UniProt database (17).

The set of amino acids that are in direct contact with the DNA was defined as all the exposed residues, with at least one atom within a distance of 6Å or less from a dsDNA atom. As expected (2), an examination of the resulting gold standard of residues in the DNA-binding surfaces showed that there is a significant enrichment of lysine and arginine in comparison to the non-binding surfaces of the same proteins. The dataset was further studied by the construction of a position-specific scoring matrix (PSSM; 18) for each of the proteins in the dataset and by analysis of the amino acids profiles (PSSM columns) of the positions that comprise the conserved patches.

59 of the 60 ML-patches found by PatchFinder included at least one residue that was in contact with dsDNA. In 42 of these cases, the overlap was >50% of the patch size. These findings demonstrate that PatchFinder is capable of locating DNA-binding sites. We also noticed that the amino acids profiles of the positions of the ML-patches in the DNA-binding proteins have distinctive properties in comparison to conserved regions in other proteins. We anticipate that taking into account this factor, as well as the patch's geometric and electrostatic properties, will facilitate the discrimination of DNA binding sites from other functional regions in proteins, and the screening for proteins that bind DNA.

Acknowledgements: We thank Dan Halperin, Yanay Ofran and Sarel J.Fleishman for helpful discussions on this work. This study was supported by a Research Career Development Award from the Israel Cancer Research Fund.

2. REFERENCES

1. Tsuchiya, Y., Kinoshita, K. and Nakamura H. 2004. Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. Proteins 55(4):885-94.
2. Stawiski, E.W., Gregoret, L.M. and Mandel-Gutfreund, Y. 2003. Annotating nucleic acid-binding function based on protein structure. J Mol Biol. 326(4):1065-79.
3. Shanahan, H.P., Garcia, M.A., Jones, S. and Thornton, J.M. 2004. Identifying DNA-binding proteins using structural motifs and the electrostatic potential. Nucleic Acids Res. 32(16):4732-41.
4. Luscombe, N.M. and Thornton, J.M. 2002. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol. 320(5):991-1009.
5. Kuznetsov, I.B., Gou, Z., Li, R. and Hwang, S. 2006. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins.64(1):19-27.
6. Jones, S. and Thornton, J.M. 2003. Protein - DNA interactions: the story so far and a new method for prediction. Comparative and Functional Genomics. 4, 428-431.
7. Nimrod, G., Glaser, F., Steinberg, D., Ben-Tal, N. and Pupko, T. In silico identification of functional regions in proteins. 2005. Bioinformatics. 21 Suppl 1:i328-i337.
8. Lichtarge, O., Bourne, H.R. and Cohen, F.E. 1996. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol. 257(2):342-58.
9. Dean, A.M. and Golding, G.B. 2000. Enzyme evolution explained (sort of). Pac Symp Biocomput. 6-17.
10. Aloy, P., Querol, E., Aviles, F.X. and Sternberg, M.J. 2001. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol. 311(2):395-408.
11. Madabushi, S., Yao, H., Marsh, M., Kristensen, D.M., Philippi, A., Sowa, M.E. and Lichtarge, O. 2002. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol. 316(1):139-54.
12. Pupko, T., Bell, R.E., Mayrose, I., Glaser, F. and Ben-Tal, N. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics. 18 Suppl 1:S71-7.
13. de Berg, M., van Kreveld, M., Overmars, M. and Schwarzkopf, O. 2000. Computational Geometry: algorithms and applications. 2nd edition. Springer. Chapter 7.
14. Liang, J., Edelsbrunner, H. and Woodward, C. 1998. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci. 7(9):1884-97.
15. Barber, C.B., Dobkin, D.P. and Huhdanpaa, H.T. 1996. The Quickhull algorithm for convex hulls. ACM Trans. on Mathematical Software. 22(4):469-483.
16. Lo Conte, L., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. 2002. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res. 30(1):264-7.
17. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., and Yeh, L.S. 2005. The Universal Protein Resource (UniProt). Nucleic Acids Res. 33(Database issue):D154-9.
18. Gribskov, M., McLachlan, A.D. and Eisenberg, D. 1987. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A. 84(13):4355-8.

Protein Function Prediction by Matching Volumetric Models of Active Sites

Thomas A. Funkhouser,*[1] Roman A. Laskowski[2], and Janet M. Thornton[2]
[1] Princeton University, Princeton, New Jersey, 08540, USA
[2] European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK
*To whom correspondence should be addressed: funk@cs.princeton.edu

1. INTRODUCTION

The goal of the proposed project is to develop an algorithm for predicting the molecular function of a protein from a structural model of its active sites. Given the 3D atomic coordinates for a novel protein and the location of a ligand binding site, we build a model of the site cavity, match it against a database of active sites models having known molecular functions, and make predictions based on the functional annotations associated with the best matches.

This general strategy is common in structural bioinformatics. The most widely used methods represent protein active sites by sets of points representing atoms, residues, pseudo-centers, templates, and/or surface critical points and match them with algorithms based on exhaustive search, geometric hashing, or association graphs. However, these methods usually rely upon a close similarity in the geometric arrangement of several key residues, and they focus on properties of surface residues rather than of cavities, and thus they can produce false positives when similar arrangements of residues are found in proximity to very different cavities.

In this paper, we represent a protein active site by a set of 3D grids describing the volumetric properties of the void inside its cavity. Given the location of an active site, we employ a knowledge-based method to build a volumetric model of the chemical and geometric properties inside its cavity and a grid matching algorithm to detect similarities to models of other sites. Our motivation is to leverage the fact that properties in the interior of an active site cavity are more likely to be functionally preserved than the properties of any individual residue or property on the protein surface.

Our method proceeds in two main steps: modeling and matching. During the modeling step, we employ an algorithm based on X-SITE [3] to analyze the 3D atomic structure of a protein and build a grid-based description its active site. During a training phase, the algorithm analyzes proteins from the Protein Data Bank (PDB) with bound ligands and stores the spatial distribution of ligand atoms for every element type (C, N, O, and P) with respect to every amino acid type in the coordinate systems defined by Singh and Thornton [7]. Then, for every test protein, those spatial distributions are resampled to build a volumetric model representing the likelihood of finding a ligand atom of each element type at every position within the active site cavity. The resulting volumetric model is stored on a regularly sampled 3D grid.

During the matching step, we compute the similarity between two active site models by finding the sum of the correlations of grids of the same element type at the optimal relative rotation. To accelerate this step, we use Fast Rotational Matching [2], a method that computes the correlation for a pair of spherical functions at all rotations in the frequency domain. Given two sets of grids, each representing the predicted spatial distribution of ligand atoms for a particular element type within an active site cavity, our method decomposes every grid into a set of concentric spherical shells and decomposes every shell into spherical harmonics. Then, the correlation between the spherical harmonic coefficients is computed for all pairs of shells and the Wigner-D-1 transform is used to map the correlations back to the space of rotations. Finally, the maximal correlation found for any rotation is used as our measure of active site similarity. This process finds correlations between two grids at all rotations in O(N4) time for grids with NxNxN resolution, whereas O(N6) would be required for a more naïve method. In practice, our implementation runs in less than a second when matching two sets of 64x64x64 grids (0.5Å resolution within a sphere of radius 16Å).

To test these volumetric modeling and matching methods for function prediction, we performed a leaveone-out classification study to predict the bound ligand type (e.g., ATP vs. NAD vs. .) of proteins found in the PDB. This task was chosen because it provides a first step towards prediction of molecular function and it is supported by enough data to perform a systematic study over a large number of proteins.

To build our test set, we scanned the PDB and selected all protein active sites with at most 3Å resolution containing at least one bound ligand having at least 20 hetero atoms. We then retained only one example within each homology family of the CATH hierarchy in order to minimize the bias due to evolutionary inter-dependence of our test set. Finally, we grouped active sites by bound ligand type and retained only the groups with at least five examples. This process yielded 105 active sites in 6 groups (ATP, FMN, BOG, NAD, FAD, and HEM). For all of the tested active sites, we constructed models of their active site cavities and matched them using the methods described in the previous section. The rank order of matches for each active site were used in a nearest neighbor classifier to predict the bound ligand type, and the percentage of correct classifications was used to evaluate the method. For comparison, we also computed the classification rate achieved with ranks computed using FASTA [5], (a sequence-alignment program), CE [6] (a structure alignment program), SCOP [4] (a structural classification), ICP [1] (a method for aligning point sets, in this case atoms within 15Å of the active site center), and random (a randomly generated rank order). We also compare to the results achieved by fast rotational matching (FRM) when given an ideal volumetric model derived from the ligands bound in the tested active sites (this test is for comparison only . it does not represent results obtainable in practice).

The table below reports the computational costs and classification rates achieved by the tested matching methods. The first result to note is that our methods for modeling and matching volumetric models of active sites (the top two rows) provide higher classification rates than the others (□61% vs. □50%). However, this improvement comes at extra storage and compute costs. The second result to note is that the fast rotational matching algorithm can achieve very high classification rates (95%) when given an ideal volumetric model of every active site (the top row). This result suggests that developing and testing better methods for modeling the volumetric properties of active site cavities may be the best way to improve our results in the near term.
From this study, we conclude that methods for matching volumetric models of active site cavities can be useful for predicting the coarse molecular function (type of bound ligand) from the structure of a protein in its bound conformation. Further study is required to determine whether similar results can be achieved with proteins in unbound conformations and/or whether these methods can be used to predict the functions of proteins for which no function is currently known.

Matching Method	Algorithm	Storage (bytes)	Match Time (sec)	Classification Rate (%)
Our method (ideal model)	FRM	106	1	95%
Our method (X-SITE model)	FRM	106	1	61%
Point alignment	ICP	103	0.1	50%
Structural classification	SCOP	101	?	50%
Structural alignment	CE	103	10	47%
Sequence alignment	FASTA	102	0.1	46%
Random	-	0	0	34%

Table 1: Comparison of compute costs and classification rates for several protein matching methods.

Acknowledgements: The authors would like to thank the NSF, BBSRC, and Leverhulme Trust for funding this project.

3. REFERENCES

1. P. J. Besl and N. D. McKay (1992). A method for registration of 3D shapes, *PAMI,* 14(2): 239-256.
2. J.A. Kovacs and W. Wriggers (2002). Fast rotational matching, *Acta Cryst.,* D58:1282-1286.
3. R.A. Laskowski, J.M. Thornton, C. Humblet, and J. Singh (1996). X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins, *J. Mol. Biol*, 259:175-201.
4. A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol*, 247:536-540.
5. W.R. Pearson (1990). Rapid and sensitive sequence comparison with FASTP and FASTA, *MethodsEnzymol,* 183:63-98.
6. I.N. Shindyalov and P.E. Bourne (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng,* 11:739-747.
7. J. Singh and J.M. Thornton (1992). *Protein Side-Chain Interactions*, Oxford University Press.

Supervised Classification of Enzyme Residue Function using Machine Learning Methods

Eunseog Youn[1], Predrag Radivojac[2], Brandon Peters[1], Charles Moad[3], Randy Heiland[3], and Sean D. Mooney[1],*
[1]Center for Computational Biology and Bioinformatics, Department of Medical and Molecular Genetics, Indiana University School of Medicine, 46202
[2] School of Informatics, Indiana University, Bloomington, 47406
[3]Scientific and Data Analysis Lab, Pervasive Technology Labs, Indiana University, 46202
*To whom correspondence should be addressed: sdmooney@iupui.edu

1. INTRODUCTION

Structural genomics projects are determining the three dimensional structure of proteins without full characterization of their function. The first conference on automated function prediction at ISMB 2005 focused on prediction of Gene Ontology terms using protein sequence and structure. Another important aspect of function prediction is identifying residues or residue environments that are important for conferring a particular function, whether known or predicted. We, along with other groups, are using sequence, structure and conservation to identify residues or structural environments that are associated with function and are using this information to apply supervised machine learning methods to predict specific residue functional roles. We have previously developed a suite of integrated tools for doing sequence and structural environment-based similarity searches against a database for rapidly identifying conserved environments and built a website and a web service around this (http://www.sblest.org/ (1)).

Our current focus is on evaluating attributes derived from the common bioinformatics tools used here, including sequence, sequence alignments (PSI-BLAST (2)), 3-D structure (FEATURE (3) and DSSP(4)), and structural environment conservation (S-BLEST (5)) for the annotation of function. We have used support vector machines to annotate homologous and non-homologous residue positions based on a specific training set of residue functions. In order to evaluate this pipeline and these attributes for automated protein annotation, we applied it to the problem of classification of catalytic residues in enzymes. We chose catalytic residues as a validating case because there are available studies to compare against and because this is a very challengingbioinformatics problem, since only about 1-2% of residues in enzymes are catalytic by the strictest definitions. When applying our method to a well-annotated set of protein structures, we were able to rank attributes for their ability to discriminate catalytic from non-catalytic residues, and those attributes which performed best were: a measure of sequence conservation, a measure of structural conservation, solvent accessibility, and residue hydrophobicity. We also found that attributes based on structural conservation were complementary to those based on sequence conservation and that they were capable of increasing predictor performance. We have added this to our website and services at http://www.sblest.org/crp/, although users are cautioned on the performance limitations of the approach.

We have compared our method with the neural network approach used by the Thornton group. They used a neural network approach to predict catalytic residues using sequence conservation, residue type and structural attributes (6). They found that attributes including sequence conservation, secondary structure, residue type and solvent accessibility were important. In our study, we assembled a diverse set of attributes based on local sequence neighborhood, 2-D and 3-D structure, and evolutionary conservation. Then, with the goal of predicting catalytic residues in unannotated proteins we developed a classification model based on SVMs. We employed ten-fold cross-validation experiments on the dataset and collected SVM prediction scores to compute the area under the curve of an ROC curve (AUC). The goal of these experiments was threefold. First, we wanted to evaluate the attributes used in our method for annotating residue functions. Second, we wanted to evaluate whether the use of structural conservation could be complementary to sequence conservation in this context. Finally, we wanted to evaluate whether catalytic residues could be predicted in novel folds. As expected, we find that our results are similar to the raw results of the Thornton group without spatial clustering (sensitivity: 65.3%; precision: 14.4%). Attribute ranking was performed on each dataset using the area under the ROC curve. To do this, we constructed an ROC for each attribute independently then ranked them according to the decreasing AUC values. Overall, in all sets we find that

the best attributes are conservation in a sequence alignment, structural conservation, solvent accessibility and residue class. Not surprisingly, the most important structural attributes are those calculated for the catalytic residue itself, while those related to its sequence and spatial neighbors were less discriminatory. The AUC was computed using the information per position score from the output of PSI-BLAST and a novel structural conservation attribute we named SCS. We find that the SCS attribute increases the AUC value (0.87) compared to the case when sequence conservation was used alone (0.84). This analysis suggests that the structural conservation based attributes can improve classification over using sequence conservation alone. Finally, we find that when trained against a fold non-redundant dataset (based on ASTRAL 40 v1.65 (7)), sensitivity drops from to 57.0% to 51.1% and precision drops from 18.5% to 17.1%. We are happy to make our training data available upon request.

In conclusion, a future for prediction of sites in both existing and new folds is possible using a set of bioinformatic attributes based on structure and conservation. We find that sequence conservation, structure conservation, residue class and solvent accessibility represent the top attributes for classification, and our attributes for structural conservation are complementary to, but not better than alone, sequence conservation. This suggests to us that improvement of structural conservation measures of a residue environment is a useful goal, perhaps using feature selection methods. As training data becomes more balanced for non-enzymatic functional classifications and more attributes for functional annotation are defined, we believe that sensitivity and precision will improve.

2. REFERENCES

1. Peters B., Moad C., Youn E., Buffington K, Heiland R. and Mooney S.D. 2006. Identification of Similar Regions of Protein Structures Using Integrated Sequence and Structure Analysis Tools. *BMC Structural Biology*. 6:4.

2. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic AcidsRes*. 25(17): 3389-3402.

3. Bagley, S., and Altman, R. B. 1995. Characterizing the microenvironments surrounding protein sites. *Protein Science* 4(4): 622-635.

4. Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577-2637.

5. Mooney, S. D., Liang, M. P., DeConde, R., and Altman, R. B. 2005. Automated characterization of structural genomics target using the Structure-Based Local Environment Search Tool (S-BLEST). *Proteins* 61(4): 741-747.

6. Gutteridge, A., Bartlett, G. J., and Thornton, J. M. 2003. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.* 330: 719-734.

7. Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S. E. 2004. The ASTRAL compendium in 2004. Nucleic Acids Res. 32: D189-D192.

KEYNOTE

Identifying Meaningful Functional Modules in the Yeast Protein-Protein Interaction Network

Shoshana J. Wodak, Jim Vlasblom and Shuye Pu,
Center for Computational Biology, Hospital for Sick Children, 555 University Avenue,
Toronto, ON Canada M5G 1X8

1. INTRODUCTION

Two recent large-scale affinity purification and mass-spectrometry studies in the yeast S. cerevisiae [1,2] produced protein-protein interaction (PPI) data of much higher quality and coverage than ever before. The identification of functional modules in these PPI network has been a major goal of these studies. Multi-protein complexes of physically interacting proteins carry out specific functions and are therefore an important category of modules. But the task of delineating complexes from PPI networks is not straightforward. The main problem stems from the fact that protein association in vivo is a dynamic process in which interaction partners change as a function of time and cellular localization. Due to the experimental protocol and the abstract nature of common network representations, information on these key aspects is however not available. Another problem arises from the fact that although the raw experimental data have vastly improved in quality and coverage, deriving reliable PPI's from these data remains a challenge due in part to methodological limitations and inherent noise in the data. We will discuss these challenges and show that the problem of deriving meaningful protein complexes from PPI networks can be tackled by using robust graph clustering techniques and mapping the results back onto the PPI network. Applying this approach to a recent PPI network derived by Collins et al. [3] from the above-mentioned studies, we predict 400 protein complexes. These encompass a record number of complexes already deposited in the MIPS & SGD databases as well as many additional ones, a good fraction of which have support in the literature. To facilitate the validation of predicted complexes, we use the software GenePro, which enables interactive display and analysis of complexes in the context of the underlying interaction network. GenePro also provides helpful interactive views of relationships between predicted (or known) complexes and various other types of functional modules such as those derived from gene expression or genetic interaction data. The presented approach and tools provide valuable insight into the organization of cellular function.

2. REFERENCES

1. Krogan NJ, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440, 637-643.
2. Gavin AC, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440, 631-636.
3. Collins S, et al. (2006) Towards a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Submitted.

Function and Interaction Prediction using multiple motif descriptors for classified domain-domain interactions and ligand binding sites.

Andreas Henschel*, Christof Winter, Wan Kyu Kim and Michael Schroeder
Biotechnological Center, TU Dresden, Tatzberg 47-51, 01307 Dresden, Germany
*To whom correspondence should be addressed: ah@biotec.tu-dresden.de

1. INTRODUCTION

Many protein sequences are still poorly annotated. Functional characterization of a protein often is improved by the identification of novel interaction partners. Here, we aim to create descriptors for all relevant sequence parts of structurally known protein-protein and protein-ligand binding sites. These binding sites are often well-conserved (1). In contrast, the rest of the surface seems to be variable (see Figure 1A) which impedes sequence similarity searches for functionally equivalent or similar proteins. Descriptors previously used for conserved domains and interface motifs are regular expressions, weight matrices and Hidden Markov Models (HMMs), covering either sequentially consecutive stretches (2-4) or full length domains (5). In particular, HMMs were successfully employed in many sequence similarity search tools (5, 6, 7).

Based on the family level of the Structural Classification of Proteins, SCOP (8), it is possible to extract and classify all domain-domain interactions found in the Protein Data Bank, PDB (9). This classification is available in the SCOPPI database (10). SCOPPI clusters similar interfaces into interface types. As pointed out by Kim and Ison, even homologous domain pairs can associate in geometrically different ways by employing different sets of residues to form interfaces (11). Consequently, the corresponding interface profiles would differ substantially which makes profile merging meaningless. However, often a number of domain-domain interactions expose striking similarities and it is desirable to collect all instances of one interface type for the calculation of the respective interface profile. We therefore compose descriptors for all interface types in SCOPPI by merging all interface profiles describing that interface type. When data for interface types is sparse, we utilize sequence data provided by HSSP (12).

Often several sequentially remote segments contribute to a binding site (exemplified in Figure 1B). To accommodate for this phenomenon, we adopt the multiple-motif approach from PRINTS (13) to represent binding sites as a collection of small HMMs for one local binding motif thus describing only the important sequence parts that form a structural feature. Each collection member gives rise to an individual sequence similarity search using the HMMer package. The P-score of the sum of the individual search result scores can be calculated using Karlin-Altschul's sum statistics for multiple high scoring sequence segments (14, formula [5]).

We compiled a comprehensive database that comprises descriptors (interface profiles) for each interface type in SCOPPI and ligand binding sites in the PDB totaling more than 3000 interface profiles. These interface profiles characterize an interaction/ligand binding site on sequence level. Hence, given a query sequence of interest, it is possible to compare it to each interface profile thus identifying possible interaction partners including ligands. Profiles for domain-domain interactions have the advantage that both interfaces can be considered. Double sided hits increase significance, i.e. given two candidate sequences, double sided hits from an interface profile pair with respective P-scores $p1$ and $p2$ yield a joint probability of $p1 \cdot p2$. Finally, Gene Ontology (16) annotations are linked to each interface profile from the original PDB entries that were used to construct this profile. The complete list of HMMs is freely available for academics upon request.

2. FIGURES

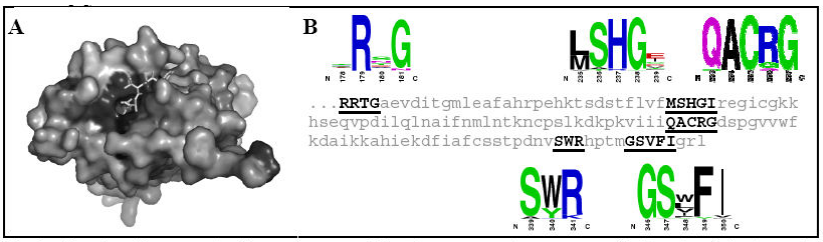


Fig.1: Constructing a set of sequence profiles to represent a conserved structural feature. A: Caspase's active site is highly conserved (1ICE, conservation levels are calculated using the von Neumann entropy and displayed in shades of gray, the darker the better conserved). Conserved residues in close vicinity of the tetrapeptide inhibitor largely define the catalytic site environment. **B:** Caspase residues within 5Å of the inhibitor are underlined. Segments are patched and those with low conservation are discarded to avoid insignificant hits. We add amino acid distribution from HSSP data for each site of the remaining segments. It is thus possible to construct HMMs and visualize the profiles as sequence logos (15).

3. REFERENCES

1. Saeed R, Deane CM. 2006. Protein protein interactions, evolutionary rate, abundance and age. *BMC Bioinformatics*. 7:128.
2. Bairoch A. 1992. PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res*. 20(Suppl): 2013–2018.
3. Espadaler J, Romero-Isart O, Jackson RM, Oliva B. 2005. Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics*. 21(16):3360-8.
4. Li H, Li J. 2005. Discovery of stable and significant binding motif pairs from PDB complexes and protein interaction datasets. *Bioinformatics*. 21(3):314-24.
5. Bateman A, Haft DH. 2002. *Brief Bioinform*. HMM-based databases in InterPro. 3(3):236-45.
6. Eddy SR: Profile hidden Markov models. *Bioinformatics* 1998, 14:755-763.
7. Zdobnov EM, Apweiler R. 2001. InterProScan-an integration platform for the signature recognition methods in InterPro. *Bioinformatics*. 17(9):847-8.
8. Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol.* 247(4):536-40.
9. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28(1):235-42.
10. Winter C, Henschel A, Kim WK, Schroeder M. 2006. SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res*. 34(Database issue):D310-4.
11. Kim WK, Ison JC. 2005. Survey of the geometric association of domain-domain interfaces. *Proteins*. 61(4):1075-88.
12. Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*. 9(1):56-68.
13. Scordis P, Flower DR, Attwood TK. 1999. FingerPRINTScan: intelligent searching of the PRINTS motif database. *Bioinformatics*. 15(10):799-806.
14. Karlin S, Altschul SF. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. Proc Natl Acad Sci U S A. 90(12):5873-7.
15. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14(6):1188-90.
16. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 25(1):25-9.

Hierarchically-Consistent Prediction of Gene Functions

T. M. Murali*
Department of Computer Science, Virginia Polytechnic Institute and State University
660 McBryde Hall, Blacksburg VA 24060
*To whom correspondence should be addressed: murali@cs.vt..edu

1. INTRODUCTION

More than 250 complete genome sequences are now available, including those of 35 eukaryotes. However, a fundamental roadblock to progress in biology is the poor state of knowledge about the biological functions of the genes in sequenced genomes [1]. Using sequence similarity to predict gene function provides annotations only for about 40% of eukaryotic genes [2]. A promising basis for predicting gene function identifies associations between pairs of genes that may perform the same or similar function in the cell. For instance, two genes may have the same function if their protein products interact or if they have very similar patterns of gene expression. A functional linkage network (FLN) is a powerful framework for representing and analysing such relationships. An FLN is a graph in which each node corresponds to a gene; the node is labelled by the set of functions that annotate the gene. An edge in an FLN connects two genes if some experimental or computational procedure suggests that these genes might share the same function. A number of powerful methods that have been published for predicting gene function by integrating different types of functional genomic data operate by constructing an FLN and then systematically propagating functional evidence across the FLN [36].

The Gene Ontology (GO) has emerged as a standard for describing the functions of gene products (where "function" is broadly defined as the molecular function of a gene product, the biological process it participates in, and the part of the cell it is localised to). The power of GO comes from the fact that it allows us to specify the function of a gene at a number of different levels of detail. In particular, the functions in GO are connected by parentchild relationships that form a directed acyclic graph (DAG). GO's true path rule specifies that if a gene is annotated with a function, then the gene must be annotated with all ancestors of that function (in the GO DAG). Function prediction algorithms can take advantage of this property by predicting functions at different levels of specificity for different genes.

However, as we explain in the full version of the paper, the functional annotations predicted by the algorithms mentioned above and other methods do not obey the true path rule. Briefly, the reasons are twofold. One, some methods focus only on a subset of functions in GO (e.g., functions at a particular depth); therefore, they cannot guarantee that predictions based on analysing the entire GO hierarchy are consistent with each other. Second, almost all methods introduce inconsistencies because they are forced to artificially generate negative examples (genefunction pairs where the gene does not have the function). We demonstrate that generating negative examples must be done carefully since the parentchild relationships in GO form a DAG and not a tree.

Our paper makes three important contributions.

1. We develop an FLNbased algorithm that provably makes predictions that follow GO's true path rule. Our approach works by modifying the approach commonly used by function prediction algorithms to generate negative examples as follows. Given a gene g and a function f, we find all other functions f' such that f' (i) does not annotate g, (ii) is a descendant of an ancestor of f, and (iii) is an ancestor of a descendant of f. For every such function f', we set the annotation status of g to be unknown, thereby allowing the possibility that we will predict f' as an annotation for g. Previous algorithms treat the pair *gf'* as a negative example, i.e., they consider *f'* not to annotate *g*. We prove that if we process the genes using the same permutation for all the functions, the Hopfieldnetworkbased algorithm used by earlier [6] will provably make hierarchically consistent predictions of gene functions.
2. We exploit this property to speed up the analysis of all functions in GO by traversing the GO DAG in topological order from root to leaf and propagating prediction results from parent to child.
3. We assign confidence estimates to our predictions by repeating our analysis with different gene permutations and counting the fraction of permutations for which we predict that a gene should be

annotated with a particular function. Our confidence estimates have an elegant monotonicity: when we predict that a gene *g* is annotated with a function *f* and with its parent *f'*, the confidence we assign to the *gf'* pair is at least as large as the confidence we assign to the *gf* pair.

Our algorithm operates correctly irrespective of the method used to construct FLNs, as long as the edge weights in the FLN remain the same for every analysed function. Further, our modified method for generating negative examples may be used by other prediction engines.

As far as we know, only one other paper [7] that deals with the issue of making hierarchically consistent predictions of gene function. These authors use independent Support Vector Machines (SVMs) for each function. These SVMs may make inconsistent predictions. The authors use a Bayesian network to combine the predictions into the probabilistically most consistent predictions. In contrast, our method mathematically guarantees the consistency of its predictions.

We apply our algorithm to an integated FLN constructed from a protein-protein interaction network for *S. cerevisiae* based on the GRID database [8] and a large compendium of gene expression profiles. We base our predictions on the GO functional annotations for *S. cerevisiae* as of March 2005. We compare our predictions with annotations in March 2006. For genes that have new annotations in March 2006, over 50% of the predictions made by our algorithm are correct. These results demonstrate that our approach makes predictions consistent with GO's true path rule and that these predictions are highly plausible and suitable for experimental followup. Upon publication of the full paper, we will make our algorithm available for use by biologists at VIRGO [9], the server we have constructed for predicting gene functions using FLNs.

2. REFERENCES

1. Roberts, R. 2004. Identifying protein functional call for community action. *PLoS Biol*, 2: E42.
2. Enright, A., Kunin, V., and Ouzounis, C. 2003. Protein families and TRIBES in genome sequence space. *Nucleic Acids Res*, 31: 4632—8.
3. Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. 2003. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol*, 21: 697700.
4. Troyanskaya, O., Dolinski, K., Owen, A., Altman, R., and Botstein, D. 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc Natl Acad Sci USA*, 100: 834853.
5. Deng, M., Tu, Z., Sun, F., and Chen, T. 2004. Mapping Gene Ontology to proteins based on protein-protein interaction data. *Bioinformatics*, 20: 895902.
6. Karaoz, U., Murali, T. M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C. R., and Kasif, S. 2004. Whole genome annotation using evidence integration in functional linkage networks. *Proc Natl Acad Sci USA,* 101: 2888—2893.
7. Barutcuoglu, Z., Schapire, R., and Troyanskaya, O. 2006. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22: 830—6. 8. Breitkreutz, B.J., Stark, C., and Tyers, M. 2003. The GRID: the General Repository for Interaction Datasets. *Genome Biol*, 4: R23.
9. Massjouni, N., Rivera, C., and Murali, T. M. 2006. VIRGO: Computational Prediction of Gene Functions. *Nucleic Acids Res,* in press.

KEYNOTE

TBA

Terry Gaasterland, Scripps Institute of Oceanography, LA Jolla, CA, USA

Classifying Protein Structures using Support Vector Machines

Jian Qiu[a], Martial Hue[c], Asa Ben-Hur[d], Jean-Philippe Vert[c], William Stafford Noble[a,b]
[a] Department of Genome Sciences, University of Washington, Seattle, WA, USA,
[b] Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA,
[c] Center for Computational Biology, Ecole des Mines de Paris, Fontainebleau, France
[d] Department of Computer Science, Colorado State University, Colorado, USA
To whom correspondence should be addressed: noble@gs.washington.edu

1. INTRODUCTION

Given the large amount of effort currently devoted to determining protein structures, the downstream problem of determining a protein's function from its structure is increasingly important. This *structure annotation problem* can be formalized as a classification task. If we assume that functional annotations are drawn from a finite list of terms, then the task consists of assigning a subset of these terms to a given structure of unknown function.

For the closely related problem of inferring protein function from primary sequence-and indeed, for a wide variety of other classification tasks in computational biology-the support vector machine (SVM) algorithm [2] has been very successful (reviewed in [5]). The SVM produces excellent classification performance due to its strong theoretical underpinnings; furthermore, the SVM's use of a generalized notion of similarity (the *kernel function*) makes the algorithm well suited to handling diverse and heterogeneous data sets.

Recently, two research groups have designed kernel functions specifically aimed at representing protein structures, with the goal of assigning functional annotations to structures. Dobson *et al.* [3] use a vector representation based upon features such as secondary structure content, amino acid propensities, surface properties, etc. In contrast, Borgwardt *et al.* [1] use a representation based upon walks defined on a graph of secondary structural elements.

In this work, we demonstrate how to construct a protein structure kernel in a straightforward fashion from an existing structural alignment algorithm, MAMMOTH [6]. We then compare the classification performance of SVMs trained using a variety of kernel functions on three distinct benchmarks: assigning enzyme classifications, SCOP superfamilies and Gene Ontology (GO) terms to protein structures. In each case, the MAMMOTH kernel performs as well or better than any of the other kernels, including the previously described vector and random walk kernels. Indeed, although the MAMMOTH kernel performs better than a kernel derived only from the amino acid sequence, some of the previously described kernels do not. Furthermore, although the SVM is capable of integrating a variety of kernels into a single classifier, we find that the combination of all the kernels yields classification performance that is no better than the MAMMOTH kernel alone.

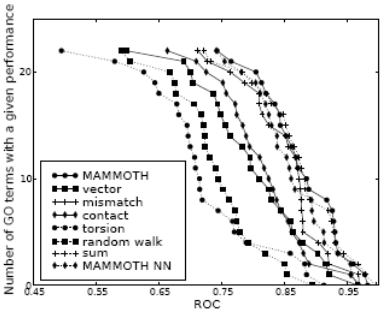
Our experiments consist of testing five protein structure kernels plus one sequence kernel on three classification benchmarks. The kernels include the vector, random walks and MAMMOTH kernels mentioned above, two additional structure kernels that rely on contact maps and C-α torsion angles, and a previously described sequence kernel (the mismatch kernel) [4]. For comparison, we also include a simple nearest-neighbor classifier using MAMMOTH E-values. The benchmarks include a previously developed enzyme classification benchmark [3], a SCOP superfamily classification benchmark, and a GO term benchmark. In this abstract we focus on the latter.

For the GO term prediction benchmark, we started with a set of 1024 PDB structures with GO annotations supported by IDA or TAS evidence codes. These structures were selected so that no two sequences share greater than 50% sequence identity. For each GO term T, we partitioned the list of proteins into three sets. First, all proteins that are annotated with T are labeled as .positive.. Next, we traverse from T along all paths to the root of the Gene Ontology graph. At each GO term along this path, we look for proteins that are assigned to that term and not to any of that term's children. We consider that such proteins might be properly assigned to T, and so we label those proteins as “uncertain”. Finally, all proteins that are not on

the path from T to the root are labeled as "negative". After this labeling procedure, we eliminated all GO terms with fewer than 30 "positive" proteins. In order to avoid redundancy, we then selected only the most specific of the remaining GO terms, i.e., the leaf nodes of the remaining hierarchy. This procedure yielded a total of 23 GO terms: 11 molecular function terms, 8 biological process terms, and 4 cellular component terms. For each term, we randomly select a subset of the negative examples, so that the ratio of negatives to positives is 3-to-1.

Classification results are summarized in Figure 1. Experiments were performed using 23 one-versus-all classification tasks, one per GO term. For each term, SVMs were tested using five-fold cross-validation, repeated three times. We measure classification performance using the area under the receiver operating characteristic (ROC) curve, and we use a Wilcoxon signed-rank test to compare methods. We observe that MAMMOTH performs significantly better than any of the other structure kernels. Furthermore, combining the MAMMOTH kernel with all of the others performs no better than using MAMMOTH by itself. Finally, we note that MAMMOTH SVM performs qualitatively better than the MAMMOTH nearest neighbor classifier, although this difference is not significant. However, on the SCOP benchmark, which is larger, the MAMMOTH SVM does perform significantly better than the nearest neighbor classifier (not shown). These results show, collectively, that an SVM with the MAMMOTH kernel is an effective way to solve the structure annotation problem.

2. FIGURES



	contact	vector	random walk	torsion
sum	0.00076	0.00076	0.00076	0.00076
MAMMOTH	0.0017	0.00099	0.00076	0.00076
mismatch	0.021	0.021	0.00076	0.00076
MAMMOTH NN	0.029	—	0.0011	0.00076
contact		—	0.0013	0.00086
vector			0.013	0.0083

Figure 1: **GO benchmark results.** The figure plots on the y-axis the number of GO terms for which a given SVM classifier achieves a specified ROC score (x-axis). Each series corresponds to a different kernel. The table lists all significant p-values for a Wilcoxon signed-rank comparison of methods. A significant p-value indicates that the method in the corresponding row performs better than the method in the corresponding column.

3. REFERENCES

1. K.M. Borgwardt, C. S. Ong, S. Schoenauer, S.V.N. Vishwanathan, A. Smola, and H-P. Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(Suppl. 1):i47.i56, 2005.
2. B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classi_ers. In D. Haussler, editor, *5th Annual ACM Workshop on COLT*, pages 144.152, Pittsburgh, PA, 1992. ACM Press.
3. P.D. Dobson and A.J. Doig. Predicting enzyme class from protein structure without alignments. *Journal of Molecular Biology*, 345:187.199, 2005.
4. C. Leslie, E. Eskin, J. Weston, and W. S. Noble. Mismatch string kernels for SVM protein classi_cation. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances in Neural Information Processing Systems*, pages 1441.1448, Cambridge, MA, 2003. MIT Press.
5. W. S. Noble. Support vector machine applications in computational biology. In J.-P. Vert B. Schoelkopf, K. Tsuda, editor, *Kernel methods in computational biology*, pages 71.92. MIT Press, Cambridge, MA, 2004.
6. A. R. Ortiz, C. E. M. Strauss, and O. Olmea. MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. *Protein Science*, 11:2606.2621, 2002.

Multi-class Protein Classification using Adaptive Codes

Eugene Ie[b], Iain Melvin[a], Rui Kuanga[c], Jason Weston[d], William Stafford Noble[e], Christina Leslie[a]
[a]Center for Computational Learning Systems, Columbia University, New York, NY,
[b]Department of Computer Science, University of California, San Diego, CA
[c]Department of Computer Science, Columbia Univerity, New York, NY
[d]NEC Labs of America, Princeton, NJ
[e]Department of Genome Sciences and Department of Computer Science and Engineering, University of Washington, Seattle, WA
To whom correspondence should be addressed: cleslie@cs.columbia.edu

1. INTRODUCTION

Predicting a protein's structural class from its amino acid sequence is a fundamental problem in computational biology and is closely linked to protein function prediction. Recently, there has been much work on novel SVM kernel methods for this problem (reviewed in [5]). Many of these methods clearly outperform standard pairwise sequence comparison algorithms and family-based generative models for the binary classification problem, i.e., discriminating between a particular protein class and all other classes. However, for these new approaches to be of practical use, kernel-based binary classifiers must be combined through some algorithmic technique into a full-fledged multi-class prediction system, where there could realistically be several hundred or over 1000 classes (protein folds or superfamilies). Clearly, it is critical to demonstrate that the accuracy of such a kernel-based multi-class prediction algorithm remains much higher than standard methods.

A general machine learning strategy for handling a multi-class prediction problem is to train a set of binary classifiers and process the binary predictions in a simple way to compute the multi-class prediction [1]. This approach assigns to each test example a vector of real-valued discriminant scores or binary prediction rule scores, which we call the output vector for the example. This class of methods includes well-known approaches such as one-vs-all, all-vs-all, and error-correcting output codes (ECOC). For example, in one-vs-all, each component of the output vector is the prediction score of a single binary classifier trained to discriminate between one class and all the others; the multi-class prediction is the class with the largest discriminant score. Though simple and widely used, the one-vs-all approach has several obvious limitations: it ignores the fact that scores of different classifiers may not be comparable, and it fails to use information about relationships between the different classes. In ECOC, one represents different classes by binary vectors or output codes in the output vector space and predicts the class based on which output code is closest to the binary output vector for the example [2].

In this work, we present a simple but effective multi-class method for protein fold and superfamily recognition that combines the predictions of binary SVM classifiers by learning in the output space. We use a state-of-the-art kernel method, called the profile kernel [4], to train the individual binary SVMs that make up the components of the output vector. In order to solve the problem that prediction scores from different classifiers are not on the same scale, we solve an optimization problem to learn a weighting of the real-valued binary classifiers that make up the components of the output code. Instead of using ad hoc output codes as in ECOC, we design codes that are directly related to the structural hierarchy of a known taxonomy, such as the manually curated Structural Classification of Proteins (SCOP), with components that correspond to fold and superfamily detectors.

Our multi-class approach involves producing a real-valued output vector $\vec{f}(x) = (f_1(x),...,f_{k+q}(x))$ for each test sequence x , where the first k component functions fi are binary SVM superfamily detectors and the next q are fold detectors, and using (k + q)-length code vectors Cj that encode the superfamily and fold of a protein class as a bit vector. We use cross-validation on the training data to learn a weight vector W = (W1, . . . ,Wk+q) to perform multi-class predictions with the weighted code prediction rule, $\hat{y} = \arg\max_j (W * \vec{f}(x))$, where $W * \vec{f}(x)$ denotes component-wise multiplication. In particular, we define a large-margin optimization problem to learn W and show that it can be solved using instances of

existing algorithms, such as the ranking perceptron or structured SVM algorithm. Figure 1 shows a summary of the adaptive code algorithm.

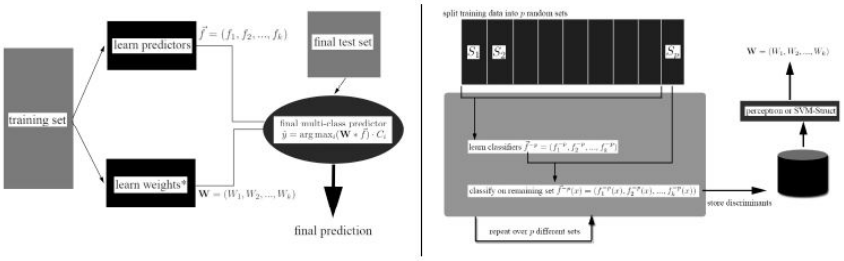


Figure 1: Summary of the main stages in our adaptive code learning method for multi-class protein classification. The black box in the left panel labeled "learn weights" is illustrated in detail in the right panel.

Method (and optimization target)	Error	Balanced Error	Top 5 Error	Balanced Top 5 Error
PSI-BLAST	0.648	0.703	0.518	0.543
one-vs-all Folds Only	0.463	0.628	0.145	0.235
one-vs-all Folds and Sfams	0.463	0.628	0.145	0.235
codes: Folds (zero-one)	0.406	0.558	0.108	0.156
codes: Folds (balanced)	0.371	0.512	0.112	0.145
codes: Folds, Sfams (zero-one)	0.409	0.552	0.117	0.172
codes: Folds, Sfams (balanced)	0.358	0.509	0.108	0.146

Table 1: Results for the fold recognition problem on a SCOP benchmark data set (26 folds, 303 superfamilies), compared to nearest neighbor using PSI-BLAST and standard one-vs-all. Error rates for both the top predicted class and the top 5 predictions are given. Zero-one error computes the error rate across all test examples, while balanced error reports the average error rate across classes.

Our code weighting approach significantly improves on the standard one-vs-all method for both the multiclass remote homology detection problem, where the test set consists of held-out SCOP protein families, and the more difficult fold recognition problem, where we hold out SCOP superfamilies. Use of codes with fold and superfamily components improves performance over fold-only codes for multi-class fold prediction in the remote homology detection setting (results not shown). Our adaptive code algorithm also strongly outperforms PSI-BLAST, used in a nearest neighbor method for multi-class prediction, on every structure classification problem we consider. Our results for the fold recognition problem are summarized in Table 1. We then extend the codes with SVM family detectors as well as probabilistic family detectors based on PSI-BLAST E-values in order to obtain significant improvement in the multi-class superfamily recognition problem, shown in Table 2. A preliminary version of this work appeared in a conference proceedings [3].

2. REFERENCES

1. Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. In Proc. 17th International Conf. on Machine Learning, pages 9–16. Morgan Kaufmann, San Francisco, CA, 2000.
2. Koby Crammer and Yoram Singer. On the learnability and design of output codes for multiclass problems. In Computational Learning Theory, pages 35–46, 2000.
3. Eugene Ie, Jason Weston, William Stafford Noble, and Christina Leslie. Multi-class protein fold recognition using adaptive codes. Proceedings of the 22nd International Conference on Machine Learning, 2005.
4. R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie. Profile-based string kernels for remote homology detection and motif extraction. In 3rd International IEEE Computer Society Computational Systems Bioinformatics Conference, pages 152–160. IEEE Computer Society, 2004.
5. W. S. Noble. Support vector machine applications in computational biology. In J.-P. Vert B. Schoelkopf, K. Tsuda, editor, Kernel methods in computational biology, pages 71–92. MIT Press, Cambridge, MA, 2004.

KEYNOTE

Novel Ways to Think About Protein Structure and its Impact on Function Prediction

Philip E. Bourne
Skaggs School of Pharmacy and Pharmaceutical Sciences,
University of California San Diego

1. INTRODUCTION

Predicting biological function from protein structure traditionally relies on a static model of the protein either implied or derived from experiment. That model is represented as a set of Cartesian coordinates. The premise of this talk is that such an approach to function prediction is limited since a protein is neither static nor are the features normally derived from the atomic coordinates the only way to think about a protein. Molecular dynamics has a long history in addressing the issue of mobility; however, it is limited as a high throughput technique. We report on recent efforts using the Gaussian Network Model (GSM) to define functionally flexible regions of a protein [1] and how they can be applied [2] on a proteomic scale. To address the issue of representation we report on recent work bridging the gap between atomic-level structure and overall protein-level functionality [3]. Such a description requires parameterization of the protein structure (and other physicochemical properties) in a quasi-continuous range, from a simple collection of unrelated amino acids coordinates to the highly synergistic organization of the whole protein entity, from a microscopic view in which each atom is completely resolved to a macroscopic description such as the one encoded in the three-dimensional protein shape. The representation uses multipoles associated with C alpha coordinates as shape descriptors.

2. REFERENCES

1. J. Gu, M. Gribskov and P.E. Bourne (2006) Wiggle – Predicting Functionally Flexible Regions from Primary Sequence PLoS Computational Biology, in revision.
2. J. Gu and P.E. Bourne (2006) Allosteric Fluctuation Transitions as Applied to Cyclin Dependent Kinase 2 JMB, Submitted.
3. A. Gramada and P.E. Bourne (2006) Multipolar Representation of Protein Structure, BMC Bioinformatics, 7:242.

A New Algorithm for the Geometrical Characterization of Protein Structures and its Application in Predicting Protein-Ligand Interactions

Lei Xie1, Philip E. Bourne1, 2,*
1San Diego Supercomputer Center, 2Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
*To whom correspondence should be addressed: bourne@sdsc.edu

1. INTRODUCTION

The 3D structure of a protein is an important asset in determining biological function(s). Those functions are characterized by specific sites on the 3-D structure through which interactions with small molecule ligands and other macromolecules take place. Identifying these sites and the partners involved in the interaction has a strong bearing on our ability to characterize structural genomics targets, study protein evolution and diversity, discover new therapeutics and engineer proteins. Not surprisingly, functional site prediction is an active area of study (1). Any high-throughput study attempting to find binding partners for a large number of proteins, as is typical of proteomic studies, needs to be efficient, robust and accurate. Here we introduce a new algorithm that is both fast and suitable for use on less well defined proteins which have been either modeled or have inherent flexibility. The algorithm is based on the notion of "geometric potential" which is analogous to hydrophobicity or electrostatics potential in that it is dependant on both the global shape of the protein structure as well as the surrounding environment of the residue (see below). Conceptually the algorithm is implemented as follows. First, the protein structure is represented by Cα atoms only, making it computationally efficient and applicable to low resolution structures and homology models on a proteome-wide scale. Second, the structure is tessellated using a convex hull algorithm (2) (http://www.qhull.org), such that protein space is partitioned into three parts, that occupied by the protein, that occupied by virtual atoms (potential ligand binding sites) and that occupied by the surrounding solvent. Hence, the partitions generate two boundaries: one surrounding the protein, and one surrounding the protein and the virtual atoms. Associated with each Cα atom are two vectors describing the relationship to each boundary (distance and direction). These vectors are used to compute the geometric potential. The value of the geometric potential depends on the residue's distance to the environmental boundary, and the distances and orientations to neighbor residues in the open space. In this way the geometric potential can be used to predict the location, boundary and significance of the ligand binding site.

The algorithm was tested on 5263 non-redundant protein-ligand complexes of both enzymes and non-enzymes where the sequence identity was below 90%. The geometric potential can distinguish ligand binding sites from non-ligand-binding site with high confidence. Figure 1 illustrates the distribution of the geometric potential for individual residues (Fig. 1a) and surface patches (Fig. 1b) with or without known ligand binding, respectively. Figure 2 illustrates that the geometric potential provides high specificity and sensitivity. Moreover, the geometric potential provides well-defined boundaries at the predicted site (Fig. 2b), a notable shortcoming of other methods (See Ben-Shimon, A. et al. (3) and references therein). When combined with other information such as evolutionary or physical properties, the geometric potential is a valuable feature in predicting functional sites. Importantly, virtual ligands can be derived from the virtual atoms that are simultaneously generated by the algorithm. This raises the possibility of identifying natural cofactors and performing fast virtual compound screening. This work is on-going and will be described. This includes a study of predicted ligand binding sites on experimental and model structures.

2. FIGURES

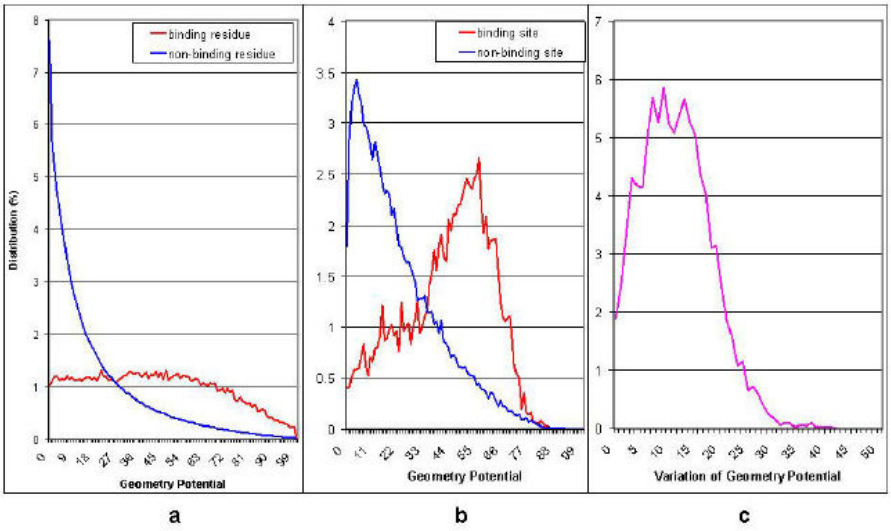


Figure 1. Geometric potential distributions of known protein-ligand complexes with and without ligand binding for: (a) single residues; (b) residue clusters that correspond to the ligand binding site and those randomly generated; and (c) the variance of the geometric potential of the binding site. The data are from 48,819 and 1,414,293 binding and non-binding residues, respectively and 7,570 and 54,826 residue clusters with and without the known ligand binding, respectively.

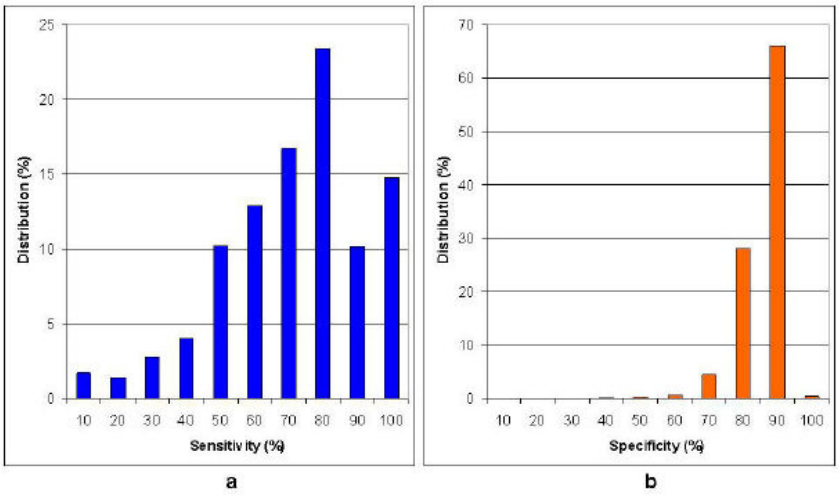


Figure 2. The (a) sensitivitiy and (b) specificity of the predicted residue clusters that overlap with the known ligand binding sites.

3. REFERENCES

1. Campbell, S. J., Gold, N. D., Jackson, R. M., Westhead, D. R. 2003 Ligand binding: functional site location, similarity and docking. Curr. Opin. Struct. Biol. 13:389-395

2. Barber, C. B., Dobkin D. P. and Huhdanpaa, H. 1996 The Quickhull algorithm for convex hulls. ACM Transactions On Mathematical Software 22:469-483.

3. Ben-Shimon, A., and Eisenstein, M. 2005. Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. J. Mol. Biol. 351:309-326.

The AnnoLite Program for Rapid and Reliable Comparative Annotation of Protein Structures

Andrea Rossi1, Fátima Al-Shahrour2, Fred P. Davis1, Ursula Pieper1,
Joaquín Dopazo2, Andrej Sali1 and Marc A. Marti-Renom1,2,*
1. Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and
California Institute for Quantitative Biomedical Research,
University of California at San Francisco, San Francisco, CA 94143, USA
2. Department of Bioinformatics, and Functional Genomics Node (INB),
Centro de Investigación Príncipe Felipe Autopista del Saler 16, 46013 Valencia, Spain
*To whom correspondence should be addressed: marcius@salilab.org

1. INTRODUCTION

Genome sequencing efforts are providing us with complete genetic blueprints for hundreds of organisms, including humans. We are now faced with assigning, understanding and modifying the functions of proteins encoded by these genomes. This task is generally facilitated by protein 3D structures, which are best determined by experimental methods such as X-ray crystallography and NMR spectroscopy. The number of known proteins structures deposited in the Protein Data Bank (PDB) has grown exponentially over the last 30 years [1]. This trend is expected to be maintained and even increased due to the Structural Genomics efforts that aim to structurally characterize most protein sequences by a combination of experiment and prediction [2, 3]. However, protein target selection by the Structural Genomics Consortia is generally not motivated by specific biological questions. Structural Genomics aims to cover the structural space by selecting targets from groups of proteins of unknown structure [4]. During recent years, more than 2,165 structures have been deposited in the PDB from Structural Genomics initiatives. About 1,900 of those contain no information about their biochemical function (*ie*, not known CATH, SCOP, InterPro, PFAM, EC or GO annotations). Moreover, this number is likely to increase with the growing output of protein structure determination techniques. Therefore, reliable and rapid methods for functional annotation of protein structures are needed to leverage the wealth of information generated by Structural Genomics.

Currently there are ~36,000 protein structure entries deposited in the PDB [1], corresponding to ~76,000 protein chains. However, this large number of structures assumes only ~700 different folds [5]. Comparative annotation benefits from two properties of the protein structures: i) the number of unique folds is far less than the number of proteins and ii) evolution tends to conserve more function and structure than sequence.

Here we introduce a simple and rapid method for functional annotation of protein structures, which transfers to the query structure known annotation of similar structures in the PDB. The method has been implemented in the AnnoLite program and relies on the DBAli [6], ModBase [7], PIBASE [8], Ligbase [9], and MSD [10] databases. Given a query structure, the AnnoLite program searches DBAli for structurally similiar proteins and collects their known annotations. Next, a p-value score is calculated for each transferred annotation using a Fisher's exact test for 2x2 contingency tables comparing two groups of annotated chains (*ie*, the group of similar chains to the query and the group of all annotated chains in the PDB)[11]. Currently, AnnoLite annotates the input protein structure with CATH [5] and SCOP [12] fold assignments, EC numbers [13], InterPro entries [14], PFAM families [15] and Gene Ontology codes [16]. The accuracy and coverage of AnnoLite were benchmarked with a set of fully annotated 1,879 nonredundant PDB chains [17]. AnnoLite can reliably annotate a structure for all of the functional properties at the exception of the GO cellular component term (Table 1). For example, the CATH fold can be recovered for 96% of the dataset with a reliability of 89% and direct functional annotation with EC numbers and Gene Ontology molecular function codes can be recovered with reliabilities of 81 and 74% for about 83 and 88% of the dataset, respectively (Table 1).

We have applied AnnoLite to all structures determined by Structural Genomics [18]. A more advanced program for comparative protein structure annotation, which also predicts domain boundaries, ligand binding sites and protein partners, has been also implemented in DBAli and will be discussed.

2. TABLES

Table 1. Accuracy and coverage of AnnoLite program. Accuracy is calculated as the ratio between correctly predicted annotations with a score smaller or equal to the cut-off score and predicted annotations. Coverage is calculated as the ratio between correctly predicted annotations with a score smaller or equal to the cut-off score and all correctly predicted annotations. Results are given in percentil.

	Optimal Cut-off	Accuracy (%)	Coverage (%)
SCOP fold	1e-4	85.3	99.0
CATH fold	1e-3	89.2	96.1
InterPro	1e-4	77.6	83.9
PFAM Family	1e-4	81.0	90.1
EC Number	1e-6	81.1	88.3
GO Molecular Function	1e-1	74.3	83.5
GO Biological Process	1e-3	72.0	85.9
GO Cellular component	1e-2	55.7	77.5

3. REFERENCES

1. Berman, H.M., et al., *The Protein Data Bank.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 6 No 1): p. 899-907.
2. Burley, S.K., et al., *Structural genomics: beyond the human genome project.* Nat.Genet., 1999. **23**(2): p. 151-157.
3. Vitkup, D., et al., *Completeness in structural genomics.* Nat Struct Biol, 2001. **8**: p. 559-566.
4. Sali, A., *100,000 protein structures for the biologist.* Nat.Struct.Biol., 1998. **5**(12): p. 1029-1032.
5. Orengo, C.A., et al., *CATH--a hierarchic classification of protein domain structures.* Structure, 1997. **5**: p. 1093-1108.
6. Marti-Renom, M.A., V.A. Ilyin, and A. Sali, *DBAli: a database of protein structure alignments.* Bioinformatics, 2001. **17**(8): p. 746-7.
7. Pieper, U., et al., *MODBASE, a database of annotated comparative protein structure models, and associated resources.* Nucleic Acids Res, 2004. **32 Database issue**: p. D217-22.
8. Davis, F.P. and A. Sali, *PIBASE: a comprehensive database of structurally defined protein interfaces.* Bioinformatics, 2005. **21**(9): p. 1901-7.
9. Stuart, A.C., V.A. Ilyin, and A. Sali, *LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures.* Bioinformatics, 2002. **18**(1): p. 200-1.
10. Boutselakis, H., et al., *E-MSD: the European Bioinformatics Institute Macromolecular Structure Database.* Nucleic Acids Res, 2003. **31**(1): p. 458-462.
11. Al-Shahrour, F., R. Diaz-Uriarte, and J. Dopazo, *FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.* Bioinformatics, 2004. **20**(4): p. 578-80.
12. Andreeva, A., et al., *SCOP database in 2004: refinements integrate structure and sequence family data.* Nucleic Acids Res, 2004. **32 Database issue**: p. D226-9.
13. Bairoch, A., *The ENZYME database in 2000.* Nucleic Acids Res, 2000. **28**(1): p. 304-5.
14. Mulder, N.J., et al., *InterPro, progress and status in 2005.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D201-5.
15. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2004. **32 Database issue**: p. D138-41.
16. Camon, E., et al., *The Gene Ontology Annotation (GOA) project: implementation of GO in SWISSPROT, TrEMBL, and InterPro.* Genome Res, 2003. **13**(4): p. 662-72.
17. http://www.salilab.org/DBAli/pages/AnnoLite_dataset.html
18. http://www.salilab.org/DBAli/pages/sg_entries.html

Voltage-gated Ion Channels and Auxiliary Subunits in the Transport Classification System

Tsai-Tien Tseng1*, Milton H. Saier Jr. 2
11776 Valley Rd., Champaign, IL 61820 USA
2University of California, San Diego, 0116 La Jolla, CA 92093-0116 USA
*To whom correspondence should be addressed: tsaitien@gmail.com

1. INTRODUCTION

The voltage-gated ion channel (VIC) superfamily is the largest and possibly the oldest of the recognized transporter protein families. Voltage-gated ion channels are known to be ubiquitous in the three kingdoms of life. The multiple topological types observed in the VIC superfamily are rare among all transmembrane transporters. It has been proposed that the architecture of voltage-gated ion channels was a result of modular design in their evolutionary history. Sequence, topological and phylogenetic analyses carried out by various labs have demonstrated strong support for this hypothesis of modularity (Nelson *et al*., 1999).

Recognized members of the VIC superfamily transport potassium, sodium and calcium ions. Most of the well characterized channels are of eukaryotic origin. Homologues of these channels are encoded in bacterial and archaeal genomes. However, calcium and sodium channels are encoded exclusively in bacterial and eukaryotic genomes. Although bacterial channels are largely uncharacterized, a recent study by Ren *et al.* concluded that several bacterial sodium channels might play a role in chemotaxis. Voltage-gated ion channels are responsible for conduction of electrical signals in excitable tissues. Excitable tissues in eukaryotes rely on the concerted coordination among various ion channels. The basic architecture for an ion channel complex involves a pore-forming principal subunit and multiple auxiliary subunits.

Phylogenetic analyses for voltage-gated ion channels and their auxiliary subunits were carried out in this report and incorporated into the Transport Classification (TC) system. The International Union of Biochemistry and Molecular Biology (IUBMB) reviewed and adopted the TC system. Unlike the Enzyme Commission (EC) system, the TC system utilizes evolution in addition to function to achieve classification. Our analysis will allow annotation to be carried out within a phylogenetic framework while utilizing the Transport Classification system (Saier, 2000). Corresponding TC numbers will allow standardized nomenclature for the annotation process.

NCBI-BLAST was used to retrieve homologues of the VIC superfamily (Altschul *et al.*, 1997). ClustalW and CINEMA5 were used for the construction of multiple alignment and phylogenetic trees (Thompson *et al*., 1997 and Pettifer, *et al*., 2004). Phylogenetic trees for sodium and calcium channels were derived using proml in the Phylip package (Felsenstein, 2004). Corresponding entries in the Transport Classification (TC) system for families described here can be accessed via the world wide web at www.tcdb.org.

There were three basic goals in our effort towards understanding ion channels in various genomes. First, a standardized classification system based on currently recognized channel proteins would be essential in assisting communication among scientists. Second, systematic analysis of genomes will elucidate the physiological importance of ion channels in a variety of organisms. Third, the deduction of sequence function relationships and modular design of ion channels needs to be studied. After overcoming the above challenges, the result of our analysis was incorporated into the Transport Classification (TC) system (www.tcdb.org) (Saier, 2000).

The work presented here described the incorporation of results from phylogenetic analyses of the principal subunits and several families of auxiliary subunits into the TC system. The VIC superfamily is currently assigned as TC# 1.A.1. The clustering pattern for calcium channels also suggested that major gene duplication events occurred prior to the separation of major animal divisions. Phylogenetic trees also lead to the suggestion that the earliest sodium channel gene diverged from an ancestral calcium channel. In addition to voltage-gated ion channels in excitable tissues from animals, this study also included ion channels of bacterial origins. Three basic groups of bacterial channels formed clusters based on the

divisions of proteobacteria, low GC gram positive and high GC gram positive. The above findings were also reflected in the TC system. Furthermore, recently discovered sperm channels were also presented. Several auxiliary subunit families were assigned individual TC numbers. The following families were incorporated into the TC system under class 8: potassium channel auxiliary subunits (Kvß [TC#8.A.5], Sloβ [TC#8.A.14], KchAP [TC#8.A.15], and MinK [TC#8.A.10]), calcium channel auxiliary subunits (α2δ [TC#8.A.18], β [TC#8.A.22], γ [TC#8.A.16]), and sodium channel auxiliary subunits (TipE [TC#8.A.19] and ß [TC#8.A.17]).

Instead of developing new algorithms for open reading frame annotation, our study focused on characterization of individual families via phylogenetic analysis and assignment of TC numbers to standardize the nomenclature. Since construction of phylogenetic trees required multiple alignments, these alignments would be readily available for extraction of motifs or position-specific scoring matrices (PSSMs). Each motif can serve as a family-specific "fingerprint." It is our hope that these family-specific motifs can then be integrated into software packages or pipelines to provide clues during automated function prediction of open reading frames (ORFs) in the near future. In essence, the analyses presented here will (a) allow family assignment for putative channels; (b) allow educated guesses on the potential physiological characteristics of putative channels; (c) serve as the main nomenclature for annotation in major databases; and (d) serve as a supplemental source of annotation for existing ion channel nomenclatures.

2. REFERENCES

1. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.

2. Nelson, R.D., Kuan, G., Saier, M.H., Jr., and Montal, M. (1999) Modular assembly of voltage-gated channel proteins: a sequence analysis and phylogenetic study. *J Mol Microbiol Biotechnol* **1**: 281-287.

3. Pettifer, S., Sinnott, J. R. , and Attwood, T. K. (2004) CINEMA - A UTOPIAn sequence editor. *CCP11 Newsletter* **Jan**.

4. Ren, D., Navarro, B., Xu, H., Yue, L., Shi, Q., and Clapham, D.E. (2001a) A prokaryotic voltage-gated sodium channel. *Science* **294**: 2372-2375.

5. Saier, M.H., Jr. (2000) A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol Mol Biol Rev* **64**: 354-411.

6. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**: 4876-4882.

Protein Motif Discovery with Particle Swarm Algorithm

Bill Chang*, Glyn Roberts, Jo-Ming Ong, Saman Halgamuge and Nalin Wickramarachchi
Bioinformatics Section, Dynamic Systems and Control Group,
DoMME, University of Melbourne, VIC 3010, Australia
*To whom correspondence should be addressed: billchc@unimelb.edu.au

1. INTRODUCTION

A protein sequence motif or a consensus pattern is a short subset of a protein sequence that constitutes a unique signature pattern identifying a family of proteins. Protein motif distinguishes members of a protein family from other unrelated proteins. Such a signature or a motif could be readily used to assign a newly sequenced protein to a specific family and thus leads to useful information about the biological function of the new protein. Particle swarm optimization was first proposed in (1) and one of its variant–the hierarchical particle swarm optimization with time varying acceleration coefficients (HPSO-TVAC) (2) has been applied for a limited set of protein families for motif discovery earlier (3). An alternative approach to HPSO-TVAC method, tuned specifically for the motif discovery problem, is introduced in this work. Using a pair-wise alignment algorithm as its heuristic, this modified HPSO-TVAC returns the optimal sequence of events (amino acid residues) within a given sequence family. The optimal intervals between residues is then determined through enumeration of the intervals occurring for the best alignments of the events found earlier. This modified approach has proved to be more computationally efficient than the previous implementation of HPSO-TVAC and tends to produce motifs with greater precision and recall.

Particle swarm optimization (PSO) is an evolutionary computing technique first introduced by Eberhart in 1995 (4). It mimics the goal seeking behaviour of social animals such as a school of fish or a flock of birds. In an optimization process, particles are initialized randomly at the start and each particle represents a possible solution. Here, a particle can represent a sequence motif such as GGT. In each iteration, particles move around the solution space by updating their position and velocity vectors after evaluating the fitness of particles. In the case of protein motif discovery problem, data are in symbolic form and need to be converted into numeric form before using particle swarm optimization. In (3), HPSO-TVAC algorithm is slightly modified in order to take account of the discrete nature of the solution space of molecular sequences. The alphabetical amino acid symbols are converted by mapping each amino acid symbol from *A* to *Y* into an integer in the range of 0 to 19. PSO algorithms depend on the fitness function to determine the best position of an individual and thereby to calculate the next generation velocities of all particles. The fitness function of an individual is evaluated by comparing it with the protein sequence and adding a score of 1 for each matched symbol and a score of zero for the unmatched symbols. In addition bonus points are added when an exact match of a sequence is found.

In this work, the overall strategy is depicted in Figure 1.

Each particle maintains a record of its position and velocity as well as the position at which its fitness was evaluated to be the greatest (i.e. its best position). Various fitness functions were tested for this investigation and the most successful of those were based on the alignment algorithms described by Shamir (5), namely, global alignment, local alignment, ends free space alignment and gap penalty.

In order to establish reliable and consistent measurements and comparisons of classification accuracy of different algorithms investigated in this paper, the true positive (TP), false positive (FP) and false negative (FN) hits are enumerated for each motif tested. A true positive hit is counted when a motif of a particular family scores a match in the same family of proteins, while a false positive hit is a match with a protein outside the family. A false negative hit is encountered when a motif does not find a match in the target family of proteins. The *precision* and *recall* are the statistical measures used by PROSITE (6) to evaluate the effectiveness of motifs generated. The two measures are combined into one accuracy measure by F-measure (7) which is defined by:

$$F = \frac{b^2 PR + PR}{b^2 P + R}$$, where P is precision ($p = \frac{TP}{TP + FP}$) and R is recall ($R = \frac{TP}{TP + FN}$),

and $b = 1$ is used in this study.

The results presented in this paper were obtained by running each pattern discovery algorithm on the sequence data of both PROSITE and Swissprot databases. The resulting patterns were then tested against all Swissprot sequences to identify matches with protein sequences. The number of true positive, false positive and false negative hits as well as the precision, recall and F-measure calculated. It is clear from Figure 2 that the modified HPSO-TVAC algorithm outperforms the original HPSO-TVAC algorithm in all cases. The improved fitness function and the use of amino acid class hierarchy patterns assisted the modified HPSO-TVAC algorithm in outperforming the original HPSO-TVAC algorithm.

2. FIGURES

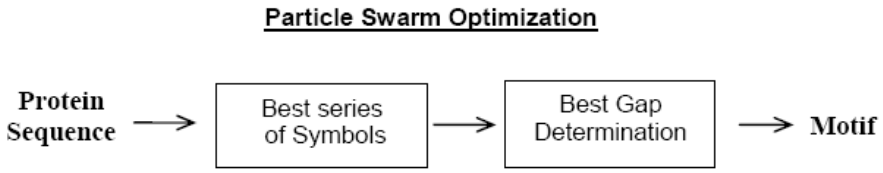


Figure 1: Modified HPSO-TVAC algorithm

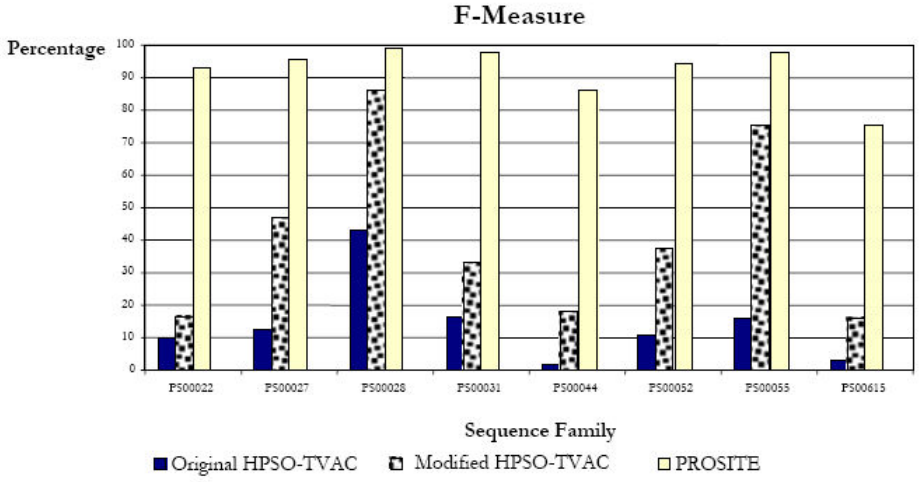


Figure 2: Comparison of Results between original HPSO-TVAC, modified HPSO-TVAC and PROSITE

3. REFERENCES

1. van Rijsbergen, C.J., Information Retrieval, London: Butterworths, 1979.

2. Ratnaweera A. C., Halgamuge S. K. and Watson H. C. (2002). Particle Swarm Optimisation with Time Varying Acceleration Coefficients. Proceedings of the International Conference on Soft Computing and Intelligent Systems 2002, Tsukuba, Japan.

3. Chang, B., Ratnaweera, A. and Halgamuge, S., "Particle swarm optimization for protein motif discovery", GENP, vol. 5 (2), 2003.

4. Eberhart, R.C. and Kennedy, J., "A new optimizer using particle swarm theory", in Proceedings of the Sixth International Symposium on Micromachine and Human Science, 1995, 39-43.

5.Shamir, R.,"Pairwise alignment" http://math.tau.ac.il/~rshamir/algmb/01/scribe02/lec02.ps.gz

6. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K. Bairoch, A., "The PROSITE Database, its status in 2002", Nucleic Acids Research, 30:235-238, 2002.
7. Smith, R. F., and Smith, T. F., "Automatic generation of primary sequence patterns from sets of related protein sequences". in Proceedings of the National Academy of Science, 1990, 118-122.

Posters

Bridging the Functional Annotation Gap with Automated Design and Matching of 3D-Templates in Protein Structures: Prediction Performance in Enzymes

David M. Kristensen[1,2], Matthew Ward[1,2], Martin Lisewski1, Brian Chen[3], Viacheslav Fofanov[4], Marek Kimmel[4], Lydia Kavraki[3,5], and Olivier Lichtarge[1,2]*

1Department of Molecular and Human Genetics, 2Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX 77030, 3Department of Computer Science, 4Department of Statistics, 5Department of Bioengineering, Rice University, Houston, Texas 77030

*To whom correspondence should be addressed: lichtarge@bcm.tmc.edu

1. INTRODUCTION

The analysis of protein function lags far behind the raw production of sequence and structure data. One approach to bridge this "annotation gap" relies on searching novel structures for 3D templates linked to specific biochemical functions. Unfortunately, such templates exist only for a fraction of the proteome, and functionally relevant matches among 3D-templates and structures are easily lost amidst random ones. Here, the ability to identify key functional residues with the Evolutionary Trace was exploited to train an automated functional annotation pipeline that picks 3D-templates and then sorts meaningful from irrelevant 3D-template matches–without any prior knowledge of catalytic sites. The positive predictive value (PPV) of the pipeline is on par with annotations by BLAST, but combining both methods raises the PPV from 48% to 62% in one test set, and from 46% to 60% in another, taken from the Protein Structure Initiative (PSI). This improvement is maintained even when considering only proteins with less than 40% sequence identity. In practice, the pipeline assigns an average of 4 or 5 putative 4-digit Enzyme Classification functions to each protein, and the true function is among them in at least 90% of cases. Further restricting function prediction to the single EC number with the most hits yields a PPV of 95% in the enzyme test set and of 77% in the PSI test set. Finally, combined with BLAST annotation, the PPV rises above 90% in both test sets. Thus, even without prior experimental knowledge of the catalytic mechanism, the function of a novel enzyme can most often be at least severely restricted if not exactly identified by combining global sequence similarity with local structural mimicry of evolutionarily important residues.

PlantAFAWE: Automatic Functional Annotation in a distributed Web Service Environment

A Joecker*, H Schoof
Max-Planck-Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany
*To whom correspondence should be addressed: joecker@mpiz-koeln.mpg.de

1. INTRODUCTION

Prediction of biological functions of proteins is an important research field. This frequently relies on translational biology, i.e. the transfer of information from a characterized protein. Additionally, a huge amount of high throughput functional genomics data is (becoming) available. This provides a data basis for intelligent filters and an automatic functional annotation. Currently, several methods for automatic protein function prediction are in use, for example homology detection, structure prediction and comparison, finding of protein domains, expression profile comparison and phylogenetic tree prediction. However, each of these methods has limited prediction accuracy. For this reason**,** the comparison of the results and decisions about cutoffs and significance of hits are task**s** of the manual annotator that are unfortunately very time-consuming.

In the last years a few hybrid approaches were published, which use machine learning algorithms for automatic prediction of GO-terms, Funcats, EC-numbers and protein classes [1][2][3]. They all have an increased accuracy, but in all cases only one feature is assigned and furthermore the tools can only handle the functional prediction of data from one organism and this data is not always up to date. Another problem is that in some cases the automatic functional prediction is not able to give a significant result. In this case combination of individually insignificant data could give clues to the function of a protein. But at the moment there is no program able to display these data.

These problems motivate to implement an automatic functional annotation system for plants in a clientserver architecture with an intuitive web interface. This will integrate multiple inputs and evaluate correlations also in available functional genomics data to improve prediction. All available analysis results can be graphically and tabularly displayed and are iteratively combined by a rule-based system based on a machine learning algorithm. In this context several different algorithms will be tested to find the most accurate one. Terms from well-known nomenclatures for protein function description (GO, EC-number, Funcat ...) and a human readable description are assigned to each protein. The user can give either a protein sequence or a keyword like a gene name or a locus code as input. The program runs analysis tools for function prediction through web services in a distributed system. This improves the scalability, accessibility, maintainability, efficiency and simplifies the process. Another advantage is that the used data is always up to date. This also means that, even in cases where the automatic machine learning algorithm does not output significant results, the tool will still be highly useful for manual annotation by summarizing in one user interface all available data on a given protein.

Web services will be executed by the Taverna Workflow Engine [4] in form of workflows (for details about the application overview see figure 1). All web services will be implemented as BioMoby web services [5] to enable automatic service discovery and strict datatyping. Relevant information will be extracted from the output and stored in a cache database to optimize performance. In case the database already contains former analysis results for a given protein sequence, this data is reused to avoid bottlenecks in data retrieval over the Internet. Cache data expires with any update of the respective sources.

All analysis tools get the protein sequence as input. If the user starts with a keyword, the protein sequence for the corresponding gene is retrieved by a web service. On the basis of this protein sequence analysis tools predict orthologous relationships between the input gene and genes in other organisms, protein domains in the input sequence, co-regulated genes in the same organism, interaction with other genes and phylogenetic relationships (see figure 2).

2. FIGURES

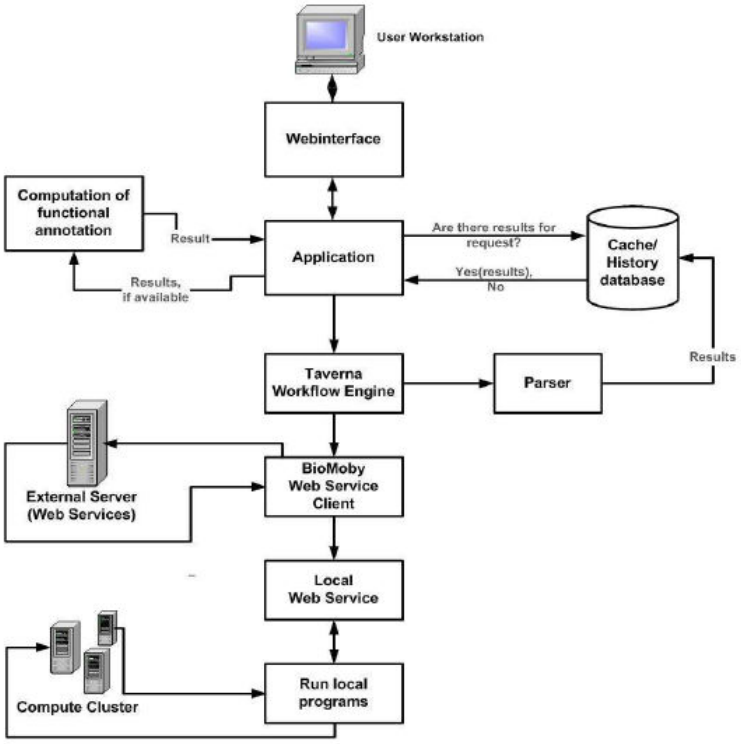


Figure 1 ***Application overview:*** *If the user has inserted a new protein sequence by using the web interface, the application will query the database for former results. If results are available, they are displayed in tabular and graphical format. If there are no results available, the application will run the Taverna Workflow Engine with predefined workflows. The Taverna Workflow Engine will run all web services which are part of the workflows. Results from all web services are stored in the cache database. New results in the cache trigger the application to run the automatic functional annotation algorithm. Available results are iteratively updated in the Web interface.*

3. REFERENCES

1. B. E. Engelhardt, M. I. Jordan, K. E. Muratore and S. E. Brenner. 2005. Protein Molecular Function Prediction by Bayesian Phylogenomics. *PLoS Computational Biology* 1(5):e45

2. Olga G. Troyanskaya, Kara Dolinski, Art B. Owen, Russ B. Altman and David Botstein. 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *PNAS* 100(14):8348-8353

3. TRUPTI JOSHI, YU CHEN, JEFFREY M. BECKER, NICKOLAI ALEXANDROV and DONG XU. 2004. Genome-Scale Gene Function Prediction Using Multiple Sources of High-Throughput Data in Yeast Saccharomyces cerevisiae. *OMICS* 8(4):322-333

4. Tom OINN et al.. 2000. Taverna: Lessons in creating a work°ow environment for the life sciences. *Concurrency Computat.: Pract. Exper.* 00:1-7

5. Mark D. Wilkinson and Matthew Links . 2002. BioMOBY: An open source biological web services proposal. *Briefings in Bioinformatics* 3(4):331-341

Evaluating Function Prediction in Mechanistically Diverse Enzymes

Alexandra M. Schnoes+, Igor Dodevski#, Patricia C. Babbitt±*
+Graduate Group in Biophysics, University of California San Francisco MC 2550, 1700 4th St., Byers Hall, 94158-2330, USA
University of Zürich Department of Biochemistry, Winterthurerstrasse 190 CH-8057 Zürich, Switzerland
±Department of Biopharmaceutical Sciences, University of California San Francisco MC 2550, 1700 4th St.,Byers Hall, 94158-2330, USA
*To whom correspondence should be addressed: babbitt@cgl.ucsf.edu

1. INTRODUCTION

Accurately predicting the functions of new uncharacterized sequences is a persistent problem for the genomic era, as is the accuracy of functional inference for proteins already annotated in public databases. We have computationally evaluated the predicted functions of proteins in several mechanistically diverse enzyme superfamilies, which pose difficult problems for prediction because their sequences, structures and conserved active sites are similar yet their overall chemical reactions, substrates and products vary widely (1). Our approach is based upon the principles of chemistry-constrained evolution, which shows that some step or characteristic of chemical activity is the conserved functional element retained in enzyme evolution rather than the ability to bind a specific substrate or catalyze an overall chemical transformation (2-4). This principle has allowed for the organization of many enzymes into their appropriate superfamilies and elaboration of principles for enzyme annotation consistent with these conserved aspects of function. Using as a gold-standard the expertly curated sequence-structure-function data available in the SFLD (Structure-Function Linkage Database (5), http://sfld.rbvi.ucsf.edu), we have designed a misannotation analysis protocol using calculated sequence thresholds and manual curation. Our evaluation of the annotations of predicted superfamily members in three commonly used public databases (NR, TrEMBL, and Swiss-Prot) and two secondary sources (Pfam, and KEGG) shows that for these types of superfamilies misannotation is a greater problem than has previously been described. We examine some of the general types of misannotation we found and discuss reasons for this growing problem. Finally, we show that even when applying expert knowledge and our misannotation protocol, some members of these superfamilies can only be annotated accurately at the superfamily level, but not at the family level, i.e. specific function cannot be assigned. Using the well-defined terminology in our SFLD ontology as a model for precise annotation and re-annotation, we propose rules for annotation aimed at predicting function only at the appropriate granularity (i.e. family level or superfamily level). We argue that this approach resolves aspects of the misannotation problem associated with over-prediction of specific function and may be generally useful for other types of proteins.

2. REFERENCES

1. Gerlt, J.A. and Babbitt, P.C. 2001. Divergent evolution of enzymatic function: Mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annual Review of Biochemistry* 70: 209-246.
2. Jensen, R.A 1976. Enzyme recruitment in evolution of new function. *Annual Review of Microbiology* 30: 409-25.
3. Petsko, G.A., Kenyon, G.L., Gerlt, J.A., Ringe, D. and Kozarich, J.W. 1993. On the origin of enzymatic species. *Trends in Biochemical Sciences* 18: 372-376.
4. Todd, A.E., Orengo, C.A, Thornton, J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology* 307: 1113-1143.
5. Pegg, S.C.-H., Brown, S., Ojha. S., Huang, C.C., Ferrin, T.E. and Babbitt, P.C. 2005. Representing structure-function relationships in mechanistically diverse enzyme superfamilies. *Pacific Symposium on Biocomputing* 2005: 358-369.

PFP: Sequence-based Annotation of Sequences and Local Sequence Motifs with Contextual GO Term Association

Troy Hawkins[1], Stan Luban[1], David La[1], Daisuke Kihara[1,2,*]
1Department of Biological Sciences
2Department of Computer Science
Purdue University, West Lafayette, IN, 47907, USA
*To whom correspondence should be addressed: dkihara@purdue.edu

1. INTRODUCTION

High-throughput techniques for experimental genomics and proteomics are driving rapid development of interpretative algorithms in bioinformatics, in particular, methods that can predict protein function using sequence, structure, gene expression, and protein-protein interaction data. Of these data, protein sequences are the most plentiful, reliable, and readily available. The incredible volume of experimentally characterized sequences and sequence motifs has allowed creation of reliable models for function annotation with far greater coverage of protein functional space than can be generated from any other single source of biological data. And although there are several existing tools which exploit the information they hold (1,2), sequences still remain a rich source of new information in bioinformatics. PFP (3) is a publicly available server (http://dragon.bio.purdue.edu/pfp/) for automated function prediction for a query sequence with Gene Ontology (GO) biological process, molecular function, and cellular component terms (4). The PFP algorithm has been shown to increase coverage of sequence-based function annotation more than fivefold by extending a PSI-BLAST search to extract and score GO terms individually and include information from distantly related sequences, and by applying a novel data mining tool, the Function Association Matrix (FAM), to score significantly associating pairs of annotations.

Based on the success of PFP at AFP-'05, we have made two major improvements: (*a*) extension of the FAM by application of association rules and $\chi 2$ significance, and (*b*) two novel applications of PFP to identify and annotate local sequence regions and functional sites.

2. MINING ANNOTATION DATABASES FOR FUNCTIONAL TERM ASSOCIATION

Using significant binary associations within sequences in UniProt to predict additional annotation terms can add 5-20% accuracy to large-scale annotation. Data mining, however, is a powerful and flexible tool that can be applied in several meaningful ways, including the clustering of predicted annotations into proteinlike" sets and automated interpretation of gene expression data. Association of two GO terms can be interpreted in three ways: (*a*) the two terms are significantly functionally coupled, i.e. they are inherently dependent on one another (at least in a particular context), e.g. membrane and transport or nucleus and transcription; (*b*) the two terms are functionally complementary, i.e. the terms are associated based on interaction of two proteins; and (*c*) the two terms are randomly associated, i.e. there is no functional significance. The first two interpretations have power in predicting terms between GO categories (67% of associations mined from UniProt) and within GO categories (33%), respectively. We applied a novel combination of $\chi 2$ significance testing (5) and general association rules (6,7) to build a database of statistically significant clusters of terms for use in PFP and other applications.

3. IDENTIFYING AND ANNOTATING FUNCTIONAL MOTIFS

We have designed two methods for identifying and annotating sequence motifs and protein domains: the first utilizes the local alignments output by the initial PSI-BLAST search to assign GO terms to distinct sequence regions (laPFP), the second runs the PFP algorithm on a sliding window of the query sequence (wPFP). The first method considers each sequence position independently; if a residue is involved in an alignment, the PFP scores previously assigned to the entire sequence are assigned to that residue additively. The result is that the signal for each term annotated to the query sequence can be viewed independently. The second method uses short windows of the query sequence to annotate narrow regions. Effectively, this method considers the conservation of the sequence of the query window. The PFP algorithm relies on

BLAST alignments for assigning scores to a query sequence, therefore those windows returning one or more alignments from a BLAST search (most return none) are considered to be more conserved and may be functionally important. In addition to finding these regions, this method also provides GO terms that can be used for their annotation. Preliminary results indicate that both are effective tools for automatically annotating functionally important subsequences in proteins of both known and unknown function (Figures 1 and 2).

4. FIGURES

4. FIGURES

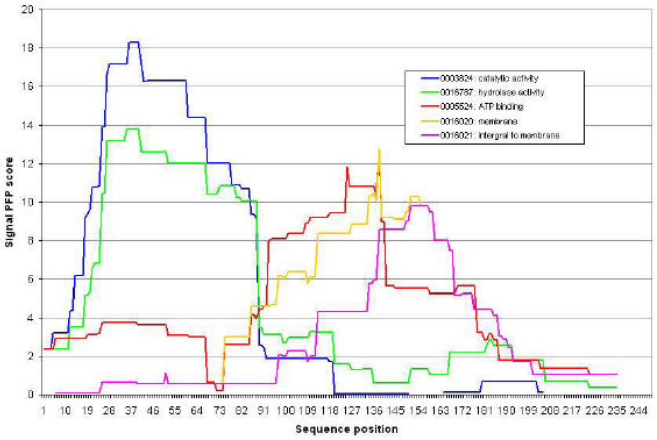


Figure 1. (*left*) laPFP assigns annotations to local sequence motifs. This example shows signal distributions obtained for a protein of unknown function. The results correctly identify two functional domains: [N-90] catalytic/hydrolase domain and [91-C] membrane-bound kinase/ATP-binding domain.

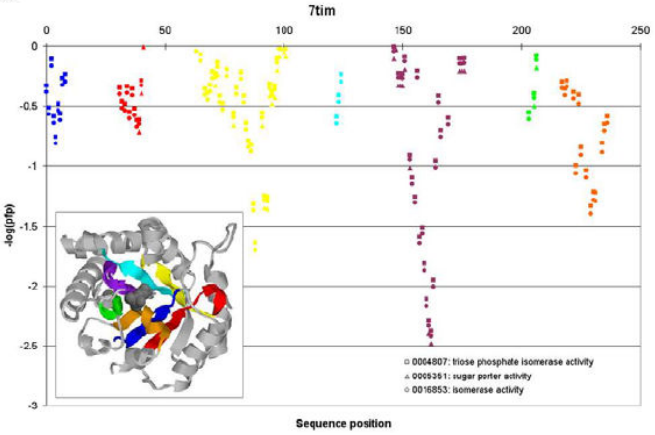


Figure 2. (*right*) wPFP identifies and annotates the 6 β-strands and loops involved in substrate binding and catalysis (colors, corresponding to inset structure) in triosephosphate isomerase (7TIM). The ligand is shown in charcoal.

5. REFERENCES

1. Pal, D. and Eisenberg, D. 2005. Inference of protein function from protein structure. *Structure (Camb.)* 13(1): 121-130.
2. Friedberg, I., Harder, T. and Godzik, A. 2006. JAFA: a Protein Function Annotation Meta Server. *Nucleic Acids Research* (in press).
3. Hawkins, T., Luban, S. and Kihara, D. 2006. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Science* 15: 1550-6.
4. Harris, M.A. *et al*. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* 32(Database issue): D258-D261.
5. Jones, C.E., Baumann, U. and Brown, A.L. 2005. Automated methods of predicting the function of biological sequences using GO and BLAST. *BMC Bioinformatics* 6: 272.
6. Toivonen, H. 1996. Sampling Large Databases for Association Rules. *VLDB 1996*: 134-45.
7. Agrawal, R., Imielinski, T. and Swami, A. 1993. Mining Associations between sets of items in massive databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data 1993*: 207-16.

Sensitive, Selective Prediction of Interaction Sites in Protein Structures

Ying Wei[1], Leonel F. Murga [1,2], and Mary Jo Ondrechen*[1]
1Department of Chemistry, Northeastern University, Boston, MA 02115 USA
2 Present address: Rosenstiel Basic Medical Sciences Center, Brandeis University, Waltham, MA 02454 USA
*To whom correspondence should be addressed: mjo@neu.edu

1. INTRODUCTION

In the present paper we establish the principle that catalytic and binding sites in proteins can be predicted from the three-dimensional structure alone with high sensitivity, good selectivity, and automated protocol. We have observed that residues involved in catalysis and recognition exhibit anomalous protonation behavior in their theoretical microscopic titration curves; we have argued that such behavior arises from strong interactions between protonation events in the vicinity of active sites and affords advantage in catalysis and in reversible binding [1-5]. Our predictions do not require any sequence or structure comparisons. Thus the method, THEMATICS, applies to novel folds and to proteins with orphan sequences. We now measure the sensitivity and selectivity of THEMATICS by comparison of its predictions to an annotated database and we also compare its predictions with those of other methods.

2. METHODS

The only input required for a THEMATICS analysis is a 3D coordinate file for the query protein in PDB format. Theoretical titration curves are computed from a Finite Difference Poisson-Boltzmann (FDPB) calculation from which their first derivative functions (which resemble probability distributions) are easily calculated by standard numerical methods. The shape anomalies of these derivative curves are well described by their third and fourth central moments that are then used to identify those cases that deviate the most from the standard Henderson-Hasselbalch shape using the statistical method of Ko *et al* [5]. Finally clusters in coordinate space of these selected residues are determined using a cutoff distance parameter. We have shown that clusters containing two or more residues occur at protein functional sites with such a high frequency, and low frequency elsewhere, that they can be considered reliable predictors of the presence of an interaction site at their location. In the present paper we show that, although the FDPB calculation is only approximate and the computed titration curve shapes are only of qualitative accuracy, the statistical analysis of the computed curves still is able to predict interaction sites with a high degree of sensitivity and good selectivity.

To measure the success of the automated method, we use a set of 168 enzymes with experimental literature annotations; this set essentially constitutes the entire original CatRes (Catalytic Residue) Database [6] (http://www.ebi.ac.uk/thornton-srv/databases/CATRES/index.html). In the present work, residues are selected using a cutoff of one standard deviation above the mean for the statistical metrics of reference [5]. Clusters of these residues are defined using a 9 Å cutoff distance.

3. RESULTS

A THEMATICS site prediction consists of a cluster in coordinate space of two or more selected residues. This prediction is considered correct if at least one residue in the cluster is annotated in the literature as an active site residue. For present purposes, we use annotations of active residues from the CatRes database and from PDBsum [7] as the reference set. For the verification set of 168 CatRes enzymes, THEMATICS shows a high sensitivity to active sites (fraction of proteins for which the correct site is identified) of 88%.

For cases where both a structure with a bound ligand and an unbound structure are available, we show that predictions are either identical or very similar for the two structures. This indicates that successful predictions can be made with unbound structures, an important capability for application to structural genomics proteins. Furthermore, the predicted cluster generally surrounds the immediate, local site where the ligand binds, another indication of the effectiveness of the method.

Selectivity is less well defined by comparison with the database because the experimental annotations are incomplete. If a residue has been shown to be part of an active site by site-directed mutagenesis and kinetics assay, for instance, then we reasonably can define it as an active site residue. However, if there is no such information in the database about a particular residue, then it is not necessarily reasonable to assume that the residue is not an active site residue. Instead, selectivity is measured by the filtration ratio, defined as the number of ionizable residues in predictive clusters divided by the total number of ionizable residues in the protein. The average filtration ratio for the 168 enzymes in the verification set is 8.2%. This low value is indicative of good selectivity and high localization.

We show that THEMATICS predictions compare very favorably with other methods. These include Patchfinder, a method that superimposes sequence conservation information onto the 3D structure [8, 9]; the structure-based methods QSiteFinder [10] and SARIG [11]; and the combined method PCats that incorporates both sequence- and structure-based information [12, 13]. It is shown that THEMATICS predictions are usually more localized, typically with lower filtration ratios, than those of Patchfinder, QSiteFinder, and SARIG. THEMATICS returns generally higher recall than PCats. However, THEMATICS and these four other methods are highly complementary. These good benchmarks indicate that the method has reliable predictive capabilities for the identification and characterization of ligand binding sites, both catalytic or recognition only. The very recent automation of the method enables high-throughput application.

Acknowledgment: NSF MCB-0517292 and the Institute for Complex Scientific Software, Northeastern University.

4. REFERENCES

1. Ondrechen, M.J., J.G. Clifton and D. Ringe, *THEMATICS: A simple computational predictor of enzyme function from structure.* Proc. Natl. Acad. Sci. (USA), 2001. **98**: p. 12473-12478.
2. Shehadi, I.A., H. Yang and M.J. Ondrechen, *Future directions in protein function prediction.* Mol Biol Reports, 2002. **29**: p. 329-335.
3. Ringe, D., Y. Wei, K.R. Boino and M.J. Ondrechen, *Protein Structure to Function: Insights from Computation.* Cellular Molecular Life Sciences, 2004. **61**: p. 387-392.
4. Murga, L.F., Y. Wei, P. Andre, J.G. Clifton, D. Ringe, and M.J. Ondrechen, *Physicochemical methods for prediction of functional information for proteins.* Israel Journal of Chemistry, 2004. **44**: p. 299-308.
5. Ko, J., L.F. Murga, P. Andre, H. Yang, M.J. Ondrechen, R.J. Williams, A. Agunwamba, and D.E. Budil, *Statistical Criteria for the Identification of Protein Active Sites Using Theoretical Microscopic Titration Curves.* Proteins: Structure Function Bioinformatics, 2005. **59**: p. 183-195.
6. Bartlett, G.J., C.T. Porter, N. Borkakoti, and J.M. Thornton, *Analysis of Catalytic Residues in Enzyme Active Sites.* J Mol Biol, 2002. **324**: p. 105-121.
7. Laskowski, R.A., V.V. Chistyakov, and J.M. Thornton, *PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids.* Nucleic Acids Res, 2005. **33**: p. D266-D268.
8. Nimrod, G., F. Glaser, D. Steinberg, N. Ben-Tal, and T. Pupko, *In silico identification of functional regions in proteins.* Bioinformatics, 2005. **21 Suppl 1**: p. i328-i337.
9. Pupko, T., R.E. Bell, I. Mayrose, F. Glaser, & N. Ben-Tal, *Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues.* Bioinformatics, 2002. **18**: p. S71-S77.
10. Laurie, A.T.R., and R.M. Jackson, *Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites.* Bioinformatics, 2005. **21**: p. 1908-1916.
11. Amitai, G., A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovski, *Newwork analysis of protein structures identifies functional residues.* J Mol Biol, 2004. **344**: p. 1135-1146.
12. Ota, M., K. Kinoshita, and K. Nishikawa, *Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation.* Journal of Molecular Biology, 2003. **327**: p. 1053-1064.
13. Kinoshita, K., and M. Ota, *P-cats: Prediction of catalytic residues in proteins from their tertiary structures.* Bioinformatics, 2005. **21**: p. 3570-3571.

A Bayesian Framework for Predicting Protein Function through Protein Interaction Networks and Homology

Richard Llewellyn and David Eisenberg*
Department of Chemistry and Biochemistry, UCLA, Los Angeles, CA 90034
*To whom correspondence should be addressed: david@mbi.ucla.edu

1. INTRODUCTION

We present a generalized probabilistic method for predicting protein function through functional-linkages (1). Our framework addresses several of the persistent problems of protein function prediction: i) how can one combine pairwise protein predictions into a single predictive distribution, ii) how can one combine homology with functional linkages, iii) how can one incorporate the strength of evidence in proteins with 'known' function into the prediction, and, iv) how can one use functional-linkages to proteins annotated with multiple disparate functions? Our method treats the function of a protein as a distribution of Gene Ontology terms (2). We integrate pairwise functional linkages by Bayesian updating of Gene Ontology biological process terms in which the likelihood is generalized through an ontological distance (3). Evidence from homology is mapped from the molecular process graph and enters as a prior distribution of biological process annotations. Our confidence in the accuracy of known functions is reflected by the precision in their annotative distributions, so that less reliable annotations receive wider distributions that have less influence on the posterior. Finally, by clustering the terms of annotated functionally-linked proteins, we return a series of predictive distributions in order to reflect the potential that unknown proteins, like experimentally-characterized proteins, may have multiple and disparate functions. We test this framework with functional-linkages identified by a spreading activation algorithm that delineates highly-connected S.cerevisea proteins from the Database of Interacting Proteins (4). Highly connected proteins, even those that have no known direct interactions, can accurately predict the function of a test set of known proteins.

2. REFERENCES

1. Marcotte, E. M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D. 1999. A combined algorithm for genome-wide prediction of protein function. Nature 402:83-86.
2. Ashburner, M. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25:25-29.
3. Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19:1275-1283.
4. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D. 2004. The Database of Interacting Proteins: 2004 update. Nucleic Acids Research 32:D449-451.

Sequence-Based Prediction of Functional Sites

Daniela Wieser, Ernst Kretschmann, Michael Kleen Rolf Apweiler*
European Bioinformatics Institute, Hinxton/Cambridge, CB10 1SD , Great Britain
*To whom correspondence should be addressed: apweiler@ebi.ac.uk

1. INTRODUCTION

Molecular functions of proteins depend on properties of their primary sequences, in particular on some site specific functional residues. We introduce a combination of the C4.5 decision tree algorithm (1) and a weighted finite state machine (WFSM) approach (2) which produces predictive models for such functional sites.

The approach for functional site prediction is depicted on figure 1 and consists of the following steps. Firstly, a decision tree is generated from a set of training proteins using taxonomic information and sequence patterns. The condition nodes of the tree represent sequence patterns identified by InterProScan and the leaf node holds the textual description of the functional site which we aim to predict (e.g. active site). Secondly, for each leaf node of the decision tree that scores above a minimal threshold, a training FSM is generated incorporating the amino acids of the training sequences. Thirdly, the training FSM is composed with an FSM that encodes mismatch penalties, gap penalties and properties of a similarity matrix (BLOSUM 62). The result encodes a predictive model which represents the training sequences and also allows some degree of mismatching. A further composition with an FSM that represents a query sequence leads to an alignment of the query sequence against the predictive model by calculating a shortest path. Finally, the predicted position of the site can be extracted from the resulting FSM. HMMs could have been used for modeling the site location. We chose WFSMs because we found them to be more generic. The method is highly efficient, does not require human interference and can be used to assign large numbers of site specific annotations to uncharacterized proteins. Taxonomic and sequence specific data are used to build the predictive models. However, the approach can be extended to include structural data as well.

The method was tested in a cross-validation against all active site annotations in UniProtKB/Swiss-Prot (3). True and false positive rates were calculated for threshold scores between 0.40 and 0.95, below which no prediction was made. The threshold score is calculated by the decision tree algorithm and is influenced by the number of true positive, true negative, false positive and false negative predictions. For a score value of 0.65, the decision tree classifier achieves a recall of over 80% at a false positive rate of 1.45% (figure 2a)while the location predictor achieves a recall value of about 90% with 3% false positive rate (figure 2b). Overall, a recall rate of 73.9% was obtained at a precision of 96.9% (graphic not shown).The prediction performance on active sites with different functional description is depicted on figure 2c.

2. FIGURES

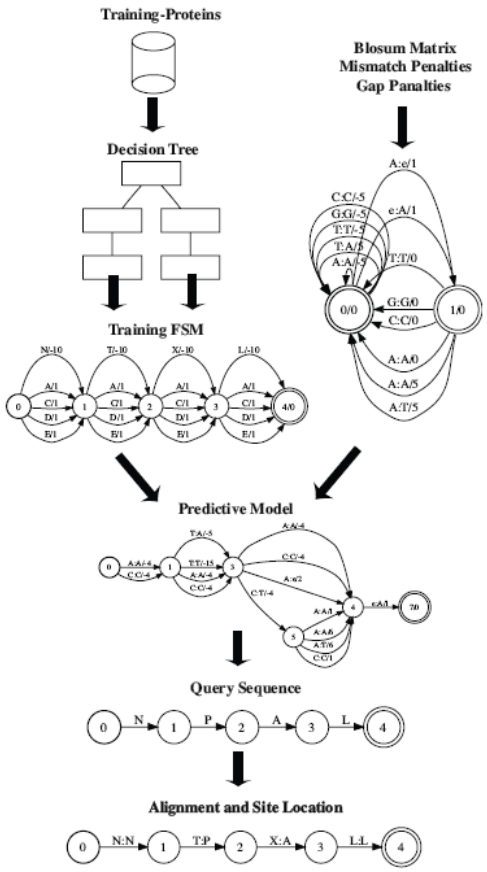


Fig.1. **Algorithm Flowchart.** Combined decision tree and weighted finite state machine approach for functional site predictions.

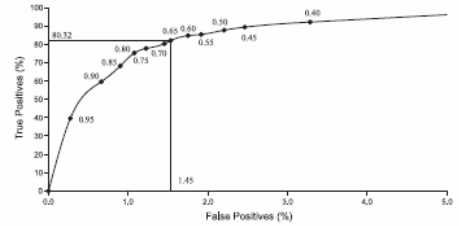


Fig.2a. **ROC Curve for Decision Tree Classifier.** The curve shows that for a score value of 0.65, a recall of 80.32% is achieved at a false positive rate of 1.45%.

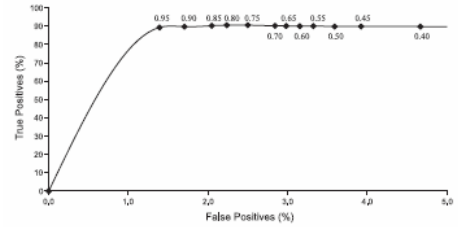


Fig.2b. **ROC Curve for Location Predictor.** The curve shows that the performance of this part of the algorithm hardly depends on the threshold score used by the binary classifier.

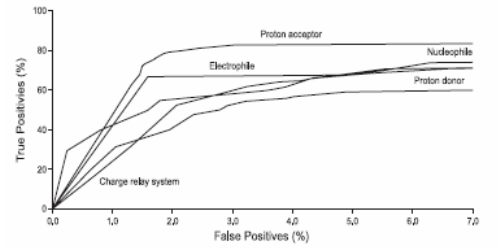


Fig.2c. **ROC Curves for Active Sites with different roles.** The curves show that some activities, like proton donor are more difficult to predict than others, like proton acceptor.

3. REFERENCES

1. Quinlan, R. 1993. *C4.5: Programs for Machine Learning*. : Morgan Kaufmann, San Francisco, CA.
2. Mohri, M., Pereira, F.C.N. and Riley, M. 2002. Weighted Finite-State Transducers in Speech Recognition. *Computer Speech and Language*, **16**, 69-88.
3. Wu, C.H. et al. 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*, **34**, D187-91.

A Novel, Local Structure-Based, Semi-Automated Algorithm for Functional Annotation of Proteomes *via* Detection of Ligand Binding Sites and Protein-Protein Interaction Partners

Vicente M. Reyes [1,3*] & Philip Bourne [1,2]
1Dept. of Pharmacology, University of California, San Diego;2San Diego Supercomputer Center, La Jolla, CA 92093-07433

*To whom correspondence should be addressed: vreyes@sdsc.edu

1. INTRODUCTION

We have developed a novel, ligand-specific, semi-automated algorithm *based on* machine learning, that can detect "signatures" of the binding sites of specific ligands in protein 3D structures, and as such can be used for protein function annotation. But in contrast to traditional machine learning methods, our algorithm detects such '3D motifs'analytically. Previously, we have shown that the algorithm has very good specificity and sensitivity for both the nucleotide ligands GTP and ATP in the small, Ras-type G-protein and ser/thrprotein kinase families, respectively. We have also previouslyscreened a library of protein structures from the proteome of *Dictyostelium discoideum*("slime mold") generated through the use of threading (123D) and side chain modeling and partial refinement (Modeller6v2), and have succeeded in assigning putative functions to some of its unannotated proteins. Experimental verification of the functional assignments using our algorithm is in progress. More recently, we have extended our algorithm to the detection of the binding site of four biologically important ligands: (a.) sialicacid, a sugar; (b.) retinoic acid, a fatty acid-like molecule; (c.) nitric oxide, a small inorganic molecule; and (d.) cadmium(II), a metal cation. We have determined the consensus 3D binding site architectures of each of these ligands from training sets consisting of experimentally determined human protein structures, and have constructed 3D search motif (the aforementioned binding site "signatures") from each in order to implement our algorithm. Our next immediate objective is to use these search motifs to screen a library of unannotated proteins whose structures have been determined experimentally through the various ongoing structural genomics programs, of which there arenow ca. 4,000 deposited in the PDB. Additionally, we have also extended our method to the prediction of protein-protein interaction partners *via*the same approach we used earlier in ligandbinding site determination. The main difference between ligand binding site detection and protein-protein interaction partner prediction is the use of pairs of search motifs in the latter, in contrast to the use of single search motifs in the former. Thus far, we have constructed search motif pairs for nine biologically important protein-protein binary complexes. As previously, our next objective will be to use these 3D search motif pairs to predict protein-protein interaction partners by screening a library of unannotated protein structures produced from ongoing structural genomics programs.

T-FUNC server for functional annotation of proteins based on structural similarity including non-sequential relations

Alexej Abyzov, Chesley Leslin and Valentin Ilyin*
Department of Biology, Northeastern University,
134 Mugar Hall, 360 Huntington Ave, Boston, MA 02143
*To whom correspondence should be addressed: ilyin@neu.edu
To whom correspondence should be addressed: vreyes@sdsc.edu

1. INTRODUCTION

The last decade in structural biology was commemorated by the rapid increase in the number of protein structures solved. Many of them have annotated functions, however often particular structural features responsible for the carried function are yet to be discovered [1]. In addition to common efforts, in recent years a Structural Genomics Initiative, aimed at solving structures with unknown functions and having low sequence similarity to known structures, has been established [2]. Up to date the initiative has resulted in 2,989 protein structures being deposited to PDB [3], with the increasing amount of deposited structures every year. Unfortunately, a significant number of those structures have little or no biological/biochemical information, including absence of functional annotation. Therefore, precise and reliable computational methods are needed to functionally annotate and analyze vast amounts of available structural data. Here we present our attempt to address this problem.

We introduce a public web-server, T-FUNC, for detailed comparative analysis and functional annotation of protein structures. The server relies on our database of structural alignments, TOPOFIT-DB (http://mozart.bio.neu.edu/topofit), containing over 80 million of structural alignments, and also a TOPOFIT one-2-all search server for comparing a user submitted structure against all known structures in the PDB. Comparative analysis of an unknown structure from the PDB can be done on the fly by retrieving data from TOPOFIT-DB database or, in case of a new structure, by calculation of structural alignments using TOPOFIT method [4] (it usually takes from 1 to 3 hours depending on the protein size). The results page displays a list of structural neighbors for the protein of interest along with links to the GO [5], SCOP [6], EC [7], and other annotation servers. The distinctive feature of the T-FUNC server is that it contains non-sequential alignments including all cases found by TOPOFIT from circular permutation to complex and completely reverse alignments. Many examples of proteins with the same function but with nonsequentially aligned regions have been reported, see for example [8, 9], and as it was found recently, nonsequential relations between proteins occur very often, at least in 1/6 cases out of all the alignments, if not more [10]. Thus, the ability to include the non-sequential alignments in the data will allow more comprehensive comparative analysis and functional annotation of a protein structure. For each alignment the user has an opportunity to visualize the corresponding alignment plot (see Figure 1), as well as to highlight the residues of interest on the plot. Structural superimposition corresponding to the alignments can be viewed in our integrated analytical front-end application, Friend [11]. The Friend software has the TOPOFIT method integrated and is capable of reproducing and visualizing alignments stored in the database.

2. FIGURES

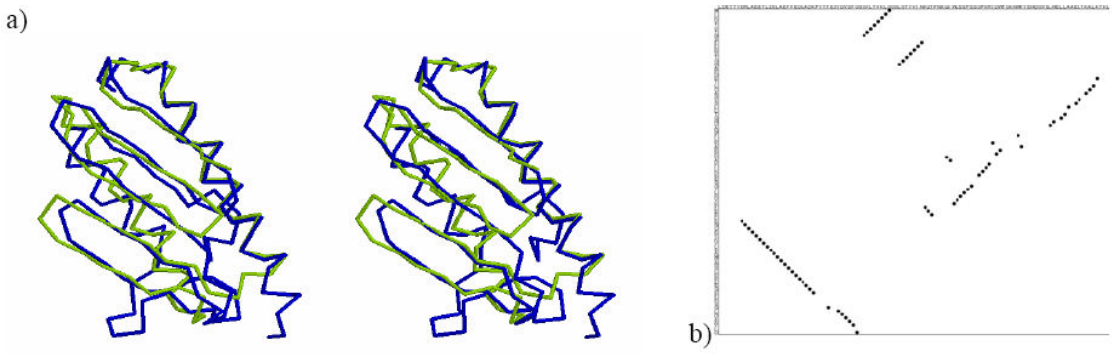


Figure 1. Structural alignment of FRATAXIN (PDB-code '1ekg' chain A) and HYPOTHETICAL PROTEIN TM1457 (PDB-code '1s12' chain A); a) stereoview of structural alignment, b) corresponding alignment plot. Structures where alignment with RMSD=1.8 Å over 74 residues. Sequence identity of aligned structures is only 17%.

3. REFERENCES

1. Laskowski, R.A., Watson, J.D., and Thornton, J.M. (2003). From protein structure to biochemical function? J Struct Funct Genomics *4*, 167-177.
2. Burley, S.K., Almo, S.C., Bonanno, J.B., Capel, M., Chance, M.R., Gaasterland, T., Lin, D., Sali, A., Studier, F.W., and Swaminathan, S. (1999). Structural genomics: beyond the human genome project. Nat Genet *23*, 151-157.
3. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.
4. Ilyin, V.A., Abyzov, A., and Leslin, C.M. (2004). Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. Protein Sci *13*, 1865-1874.
5. Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Muldr, N., Oinn, T., Maslen, J., Cox, A., and Apweiler, R. (2003). The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Res *13*, 662-672.
6. Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol *247*, 536-540.
7. Bairoch, A. (2000). The ENZYME database in 2000. Nucleic Acids Res *28*, 304-305.
8. Grishin, N.V. (2001). Fold change in evolution of protein structures. J Struct Biol *134*, 167-185.
9. Nagano, N., Orengo, C.A., and Thornton, J.M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol *321*, 741-765.
10. Abyzov, A., and Ilyin, V.A. (2006). A comprehensive analysis of non-sequential alignments between protein structures. Submitted.
11. Abyzov, A., Errami, M., Leslin, C.M., and Ilyin, V.A. (2005). Friend, an integrated analytical front-end application for bioinformatics. Bioinformatics *21*, 3677-3678.

TACT: Transcriptome Auto-annotation Conducting Tool of H-InvDB

Chisato Yamasaki[1,2], Hiroaki Kawashima[1, 3], Fusano Todokoro[3], Yasuhiro Imamizu[3], Makoto Ogawa[3], Motohiko Tanino[1, 2], Takeshi Itoh[2, 4], Takashi Gojobori[2, 5, 6] and Tadashi Imanishi*[2]

1, Integrated Database Group, Japan Biological Information Research Center, Japan Biological Informatics Consortium, AIST Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan; 2, Integrated Database Group, Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, AIST Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan; 3, DYNACOM Co., Ltd., 643 Mobara, Mobara-shi, Chiba 297-0026, Japan; 4, Genome Research Department , National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan; 5, Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan; 6, Department of Genetics, The Graduate University for Advanced Studies, 1111 Yata, Mishima. Shizuoka 411-8540, Japan.

*To whom correspondence: imanishi@jbirc.aist.go.jp

1. INTRODUCTION

Transcriptome Auto-annotation Conducting Tool (TACT) is a newly developed web-based automated tool for conducting functional annotation of transcripts by the integration of sequence similarity searches and functional motif predictions. We developed the TACT system by integrating two kinds of similarity searches, FASTY[1] and BLASTX [2], against protein sequence databases, UniProtKB (Swiss-Prot/TrEMBL) and RefSeq, and a unified motif prediction program, InterProScan[3], into the ORF-prediction pipeline originally designed for the "H-Invitational" human transcriptome annotation project[4, 5] (Fig. 1). This system successively applies these constituent programs to an mRNA sequence in order to predict the most plausible ORF and the function of the protein encoded. In this study, we applied the TACT system to 19,574 non-redundant human transcripts registered in H-InvDB, and evaluated its predictive power by the degree of agreement with human-curated functional annotation in H-InvDB. As a result, the TACT system could assign functional description to 12,559 transcripts (64.2%), the remaining being hypothetical proteins. Furthermore, the overall agreement of functional annotation with H-InvDB, including those annotated as hypothetical proteins, was 83.9% (16,432 / 19,574). These results show that the TACT system is useful for functional annotation and that the prediction of ORFs and protein functions is highly accurate and close to the results of human curation.

TACT is freely available at http://www.jbirc.aist.go.jp/tact/.

2. FIGURES

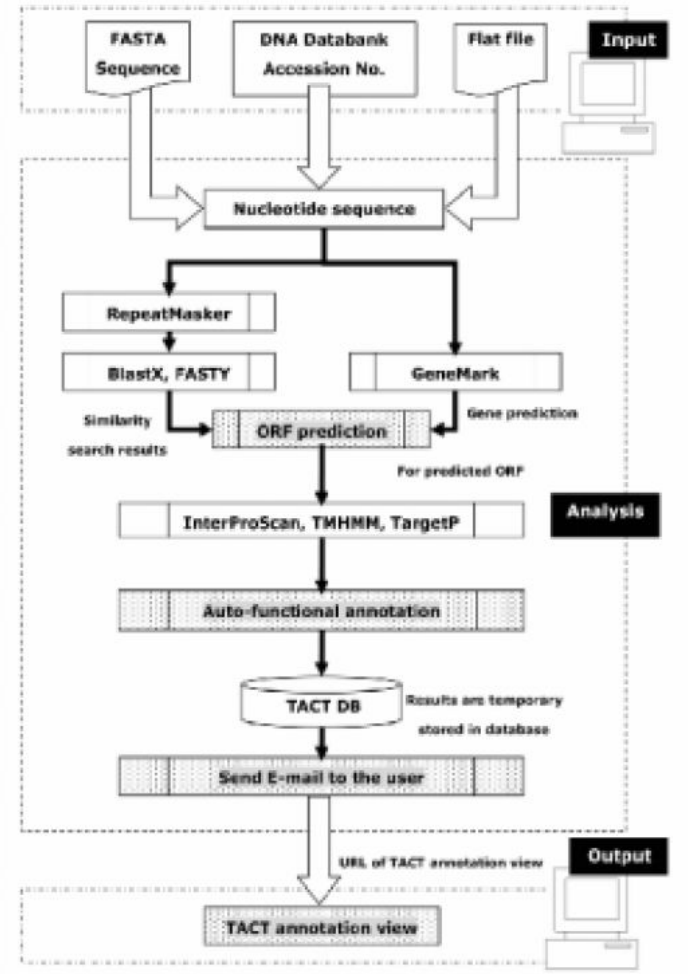


Fig 1. **The TACT annotation pipeline.**
The flowchart illustrates the TACT computational analysis and web-server interfaces. The white arrows indicate the input sequence data to TACT and output annotation data from TACT to users. The thick solid arrows indicate the data flow within the TACT server during analysis.

3. REFERENCES

1. Pearson, W.R., et al., *Comparison of DNA sequences with protein sequences.* Genomics, 1997. **46**(1): p. 24-36.
2. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
3. Mulder, N.J., et al., *InterPro, progress and status in 2005.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D201-5.
4. Imanishi, T., et al., *Integrative annotation of 21,037 human genes validated by full-length cDNA clones.* PLoS Biol, 2004. **2**(6): p. 856-75.
5. Yamasaki, C., et al., *Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB).* Gene, 2005. **364**: p. 99-107.

Assessing the contribution of functional linkages to function prediction

Arturo Medrano-Soto, Debnath Pal, David Eisenberg*
UCLA-DOE Institute for Genomics and Proteomics. 611 C. E. Young Drive East. Los Angeles, Califiornia. 90095 USA
*To whom correspondence should be addressed: david@mbi.ucla.edu

1. INTRODUCTION

The current deluge of genomic sequences, 3D structures, transcriptomic, proteomic, and protein-protein interaction data has boosted the development of a variety of computational tools in order to approach one of the foremost challenges in bioinformatics, namely automated function prediction. As previously reported, we developed a method to integrate multiple bioinformatics resources and formally weight the functional annotations obtained from them in a well-organized and coherent output. This method is implemented and accessible online through the ProKnow [1] metaserver (http://www.doe-mbi.ucla.edu/Services/ProKnow/).

In this work we investigate the extent to which the quality of function prediction can be improved by additionally taking into account GO annotations collected from functionally linked genes. Obviously, not all functionally linked genes share the same molecular function, however, we do expect that most of them will share the same biological process. Therefore, the goal is to identify the subset of functionally related genes that share similar GO annotations, which in turn will be used as additional clues for predicting gene function.

As a preliminary analysis, we studied the similarities in GO annotations for all pairs of genes in *Escherichia coli* K12 and *Bacillus subtilis* for which there is experimental evidence indicating that they reside within the same operon. We observed that close to 60% and 40% of all genes within operons have very similar biological process and molecular function GO annotations, respectively, as quantified by the method of Lord et al [2]; This is significantly higher than the similarities observed for random gene pairs (less than 3%).

Encouraged by these results, we obtained the set of nonredundant complete genomes using a previously reported method [3], and extracted from ProLinks [4] all functional linkages inferred by the Gene Cluster, Gene Neighbor, Rosetta Stone and Phylogenetic Profiles methods. We only considered predicted functionally related genes showing at least 0.65 confidence values and at most 0.05 GO distance —Cellular component GOs were not taken into account in this analysis. We finally incorporated this subset of genes in ProKnow as the ProLinks feature extractor and compared predictions when ProLinks is included and excluded from the analysis. We are currently assessing the benefit of including functional linkages in ProKnow predictions.

2. REFERENCES

1. Pal, D. and D. Eisenberg, *Inference of protein function from protein structure.* Structure, 2005. **13**(1): p. 121-30.
2. Lord, P.W., et al., *Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.* Bioinformatics, 2003. **19**(10): p. 1275-83.
3. Moreno-Hagelsieb, G. and J. Collado-Vides, *Operon conservation from the point of view of Escherichia coli, and inference of functional interdependence of gene products from genome context.* In Silico Biol, 2002. **2**(2): p. 87-95.
4. Bowers, P.M., et al., *Prolinks: a database of protein functional linkages derived from coevolution.* Genome Biol, 2004. **5**(5): p. R35.

Boosting Our Understanding of DNA-Binding Proteins

Robert Langlois and Hui Lu*
Department of Bioengineering, University of Illinois at Chicago, Chicago IL, 60607, USA
*To whom correspondence should be addressed: huilu@uic.edu

1. INTRODUCTION

At the heart of protein-function annotation lies in the complex interactions between groups of atoms forming variety of molecules from amino acids to solvent. By incorporating prior information about such groups of molecules, we apply state of the art machine learning algorithms to annotate the function of larger polymers, i.e. proteins, by identifying their interaction with other polymers. Protein-DNA interaction is one such key bimolecular interaction. DNA-binding proteins play a pivotal role in various intra- and extra-cellular activities ranging from DNA replication to control of gene expression. Recently our lab has focused on incorporating the protein in the form of numerical features that encode structure and sequence-based characteristics; then, we generate a set of rules (Fig. 1) using a supervised machine learning algorithm such as SVMs trained over a set of known cases. The resulting accuracy, 86%, outperformed all other published data [1]. In this current work, we want to address two key issues: Can we increase the performance? And what are the important features and relationships between features in discriminating DNA-binding proteins from other proteins? The dataset and related information can be found on our website at http://proteomics.bioengr.uic.edu/pro-dna/. To achieve these goals, we established a new machine learning workbench, MALIBU. The workbench includes various machine learning protocols incorporating existing software (like LIBSVM) to home-made programs using other algorithms such as decision trees, boosting, etc. The advantages of our workbench include direct comparison of various methods, a set of robust and well tested algorithms, and core set of algorithms to extend the basic classifier (e.g. isotonic calibration). For DNA-binding protein prediction, we focused on a novel implementation of AdaBoost on Decision Trees. The simplicity and power of boosted decision trees lends particular advantage over other classifiers (such as Neural Nets and SVMs) especially in efficiency. We are particularly interested in the model generated by boosting in that it has the nice feature allowing one to trade accuracy for complexity. Expanding on previous work [1, 2], we find boosted trees perform much better than SVMs over the DNA binding protein dataset [1] achieving a 2-fold iterated cross-validation accuracy of 89% and a leave-one-out cross-validation accuracy of 92% (Table 1, 2). Moreover, our analysis reveals that structural features previously thought important didn't provide too much improvement in the prediction (Fig. 1); specifically AdaBoost using sequence based features can achieve 90% testing accuracy using leave-one-out cross-validation as compared to 92% over the full set of features (Table 1). To this end, we incorporate an expanded set of sequence based features aimed at increasing the performance of annotation. While this approach is applied to identifying DNA-binding proteins, it can be extended to the prediction of other protein functions [3].

2. TABLES

Table 1: Comparing the performance of tree classifiers on feature subsets

Dataset	# Attrs.	J48	AdaTree	AdaStump
Full	42	82.3%	92.2%	91.0%
CH + AA	21	83.5%	90.4%	89.4%
AA	20	81.4%	88.2%	87.9%
PS + sAA	21	78.9%	85.1%	83.2%
sAA	20	77.6%	85.1%	82.0%

Abbreviations for the above table include: AA for amino acid composition, CH for charge, sAA for surface amino acid composition, # Attrs for number of attributes and PS for patch size. Percentages are the accuracy achieved with corresponding features.

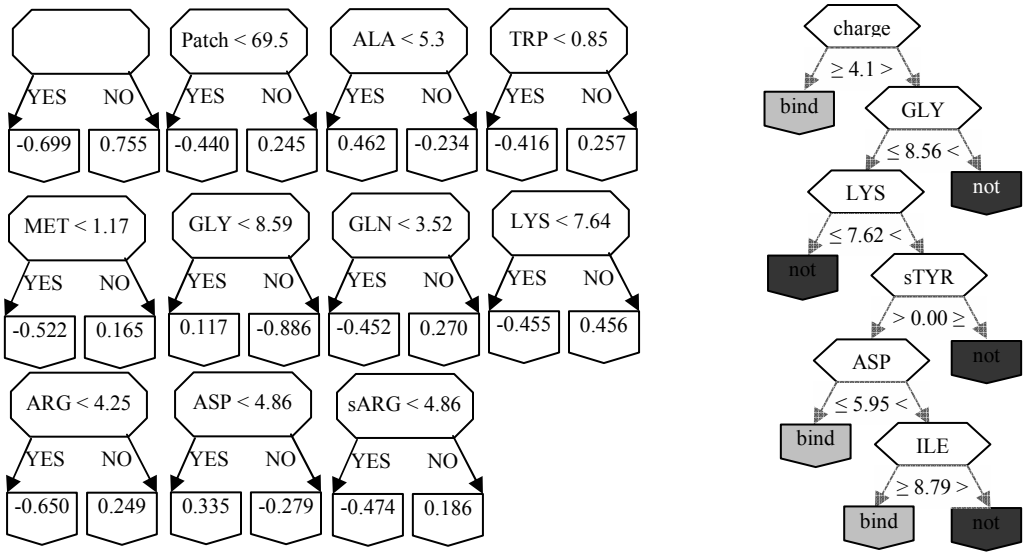
Table 2: Evaluating tree classifiers over different datasets/methods

		Full Dataset (121/238)				20% Dataset (84/238)			
Classifiers	CV	Acc.	Sen.	Spe.	AUC	Acc.	Sen.	Spe.	AUC
AdaTree	2-fold	87.1	75.2	93.2	90.8	88.5	69.1	95.4	89.6
	5-fold	89.8	80.6	94.5	93.5	90.7	73.8	96.6	91.7
	LOO	90.3	81.8	94.5	94.9	92.2	76.2	97.9	92.8
AdaStump	2-fold	84.4	74.0	89.7	88.8	86.4	67.8	93.0	88.0
	5-fold	85.6	76.8	90.1	90.4	88.6	72.1	94.4	89.7
	LOO	86.1	77.7	90.3	91.0	91.0	77.4	95.8	89.4
J48	2-fold	80.9	60.2	91.5	76.5	82.0	52.2	92.5	73.5
	5-fold	81.8	60.8	92.4	78.5	82.9	54.3	93.0	75.2
	LOO	82.2	61.2	92.9	77.9	82.3	56.0	91.6	76.4

Cross-validation was used to evaluate the classifiers averaging 2-fold over 100 runs and 5-fold over 40 runs. The first two boosting classifiers were set to the optimal number of iterations using 5-fold cross-validation over the training set. The J48 algorithm was run with default arguments. The full dataset is <35% sequecen identity; 20% dataset is <20% sequence identity. We have found the prediction models are not relying on sequence homology when comparing the performances of full set and 20% set.

3. FIGURES

Figure 1: Truncated Boosted Stump Model and Pruned Decision Tree



Illustrated are two prediction models built over the DNA-binding protein dataset. The prefixing *s* on the three letter amino acid codes stands for surface. See [1] for more details of the dataset and features.

4. REFERENCES

[1] Bhardwaj, N., Langlois, R., Zhao, G., and Lu, H. (2005) Kernel-based machine learning protocol for recognizing DNA-binding proteins. Nucleic Acids Research. 33(20): 6486-6493.

[2] Bhardwaj, N., Langlois, R., Zhao, G., and Lu, H. (2005) Structure Based Prediction of Binding Residues on DNA-binding Proteins. Proceedings of 27th IEEE EMBS Annual International Conference.

[3] Bhardwaj, N., Stahelin, R.V., Langlois, R.E., Cho, W., and Lu, H. (2006) Structural Bioinformatics Prediction of Membrane-Binding Proteins. Journal of Molecular Biology. (available online March 30, 2006)

A Bayesian Framework for Predicting Protein Function through Protein Interaction Networks and Homology

Richard Llewellyn and David Eisenberg*
Department of Chemistry and Biochemistry, UCLA, Los Angeles, CA 90034
*To whom correspondence should be addressed: david@mbi.ucla.edu

1. INTRODUCTION

We present a generalized probabilistic method for predicting protein function through functional-linkages (1). Our framework addresses several of the persistent problems of protein function prediction: i) how can one combine pairwise protein predictions into a single predictive distribution, ii) how can one combine homology with functional linkages, iii) how can one incorporate the strength of evidence in proteins with 'known' function into the prediction, and, iv) how can one use functional-linkages to proteins annotated with more than one function? Our method treats the function of a protein as a distribution of Gene Ontology terms (2). We integrate pairwise functional linkages by Bayesian updating of Gene Ontology biological process terms in which the likelihood is generalized through an ontological distance (3). Evidence from homology is mapped from the molecular process graph and enters as a prior distribution of biological process annotations. Our confidence in the accuracy of known functions is reflected by the precision in their annotative distributions, so that less reliable annotations receive wider distributions that have less influence on the posterior. Finally, by clustering the terms of annotated functionally-linked proteins, we return a series of predictive distributions in order to reflect the potential that unknown proteins, like experimentally-characterized proteins, may have multiple and disparate functions. We test this framework with functional-linkages identified by a spreading activation algorithm that delineates highly-connected *S.cerevisea* proteins from the Database of Interacting Proteins (4). Highly connected proteins, even those that have no known direct interactions, can often predict the function of a test set of known proteins.

2. REFERENCES

1. Marcotte, E. M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* 402:83-86.
2. Ashburner, M. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25:25-29.
3. Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19:1275-1283.
4. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research* 32:D449-451.

Learning Rules for Microbial Genome Annotation

Lucie Gentils1, Jérôme Azé[1], Philippe Bessiéres [2], Jean-François Gibrat [2]*, Valentin Loux [2], Céline Rouveirol1 et Christine Froidevaux [1]
1 Laboratoire de Recherche en Informatique, UMR CNRS 8623, Université Paris-Sud, 91405 Orsay, France
2 Mathématique, Informatique et Génome, INRA, Domaine de Vilvert, 78352 Jouy-en-Josas Cedex, France
*To whom correspondence should be addressed: jean-francois.gibrat@jouy.inra.fr

1. INTRODUCTION

The number of sequenced prokaryote genomes is increasing exponentially (currently there are 343 published genomes and more than 1 000 ongoing sequencing projects). Ten years ago, the first prokaryote genomes were the result of a joint effort of a consortium of laboratories. Recently an increasing number of small teams of biologists have been able to carry out the sequencing of their favorite organism.

To help these teams in the analysis of their sequenced genome we have developed an annotation platform called AGMIAL [1]. The role of this platform is to allow human experts to concentrate exclusively on the annotation, hiding the technical details, making the system implementation transparent, centralizing and facilitating the access to relevant data, and reporting a synthesis of all the findings to the annotators in an efficient manner. Yet, in order to obtain a good quality annotation, it is crucial that human experts validate the annotation, i.e., assess the results, check their consistency, possibly integrate them with data stored in databases or knowledge coming from the literature, etc. The latter stage is very time-consuming, especially for small teams, and constitutes the bottle-neck of the annotation process. Therefore the question arises: can the productivity of annotators be improved during this stage?

To address this issue we propose a semi-automatic system which will help annotators by suggesting them annotations. These ones are induced by rules that reflect known protein annotations and take into account annotators strategies. One of the main objectives of this system is to obtain rules that can be easily understood and used by human annotators. We chose a logic based framework for representing the rules: Inductive Logic Programming (ILP). ILP is a subfield of Machine Learning that focuses on rules or concepts that can be represented in first-order logic, which is fully adapted to represent complex relations as those we expect to be useful for genome annotation.

Our learning set consists of two genomes, Lactobacillus sakei [2] and Lactobacillus bulgaricus [3], that have been manually annotated by experts using the Subtilist functional hierarchy [4]. The AGMIAL platform provides, for the proteins encoded by these genomes, different pieces of information that can be broadly divided into two categories:

- intrinsic properties of the proteins, e.g., isoelectric point, length, molecular weight, presence of a particular domain, of a particular feature such as coiled-coil, low-complexity or transmembrane regions, etc.

- relations between proteins such as the relations of homology or the relations deduced from an analysis of the genomic context. Notice that these relations link together proteins of the genome under study to proteins stored in databases (such as UniprotKB) that have an annotation, e.g., GO terms, SwissProt keywords, etc.

The rules are represented by decision-trees and are learned with TILDE [5], a relational learning system from the ILP community. In this study, for the sake of testing the methodology, we only consider the homology relationship between proteins together with the associated GO terms. For the purpose of high biological significance, we choose to consider only homology where $e - value < 10^{-4}$ and $similarity - degree > 50\%$. Also, as a first approximation, we only use the first level of the functional hierarchy, which contains four classes. The rules are learned on one of the genomes and applied to the other one. Then we compare the results achieved, by the set of rules inferred by TILDE with the expert's annotation, on the second genome.

On average, for the first level of the Subtilist functional hierarchy, we obtain a precision ≥85%, a recall ≥70% and an accuracy ≥85%. On the test sets, precision is higher than 70%, recall ≥50% and accuracy ≥80%.

Here is an example of rules obtained from *L. sakei* genome. For this experiment, we choose to learn the *class3* of the Subtilist hierarchy against the three other ones. Proteins belonging to *class3* can be annotated by at least one of the following terms: "Information pathways, DNA replication, DNA restriction and modification, DNA recombination and repair, DNA packaging and segregation, RNA synthesis, RNA modification, Protein synthesis, Protein modification, Protein folding, Protein degradation".

If *the protein A has more than 4% of homologues described by the GO term "protein biosynthesis"*

then *it belongs to the class3*

otherwise If *the protein A has more than 7% of homologues described by the GO term "DNA repair"*

then *it belongs to the class3 ...*

For more rules and/or further information about these rules, see www.lri.fr/RAFALE/jobim2006/. Although more work is clearly needed for improvement, our preliminary experiments show promising results. First of all the methodology will be tested on the different levels of the Subtilist hierarchy, down to the leaves that provide a more detailed description of the function. Also all the relevant information provided by the AGMIAL platform will be taken into account. Finally we plan to assess the impact, upon the performances of the method, of the phylogenetic distance between organisms that are used to learn the rules and organisms on which these rules are applied.

2. REFERENCES

[1] K. Bryson, V. Loux, R. Bossy, P. Nicolas, S. Chaillou, M. van de Guchte, S. Penaud, E. Maguin, M. Hoebeke, P. Bessi`eres, and J-F Gibrat. AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Res, 2006. in press.

[2] S. Chaillou, M.-C. Champomier-Verg`es, M. Cornet, A.-M. Crutz-Le Coq, A.-M. Dudez, V. Martin, S. Beaufils, E. Darbon-Rong`ere, R. Bossy, V. Loux, and M. Zagorec. The complete genome sequence of the meat-borne lactic acid bacterium Lactobacillus sakei 23k. Nature Biotechnology, 23:1527–33, 2005.

[3] M. van de Guchte, S. Penaud, C. Grimaldi, V. Barbe, K. Bryson, P. Nicolas, C. Robert, S. Oztas, S. Mangenot, A. Couloux, V. Loux, R. Dervyn, R. Bossy, A. Bolotin, J-M Batto, T.Walunas, J.-F. Gibrat, P. Bessi`eres, J. Weissenbach, S.D. Ehrlich, and E. Maguin. Complete genome sequence of Lactobacillus bulgaricus: evolution caught en route. Proc Natl Acad Sci, USA, 103:9274-9279, 2006.

[4] I. Moszer, L.M. Jones, S. Moreira, C. Fabry, and A. Danchin. Subtilist : the reference database for the Bacillus subtilis genome. Nucleic Acids Res, 30:62–5, 2002.

[5] H. Blockeel and L. De Raedt. Top-down induction of first order logical decision trees. Artificial Intelligence, 101(1-2):285–297, 1998.

Recognizing Complex Ligand Binding Sites Using Multiple Local Structural Models

Jessica Ebert and Russ Altman*
Department of Genetics, Stanford University, 300 Pasteur Drive, Lane L301, Stanford, CA 94305, USA
*To whom correspondence should be addressed: russ.altman@stanford.edu

1. INTRODUCTION

While most functional prediction efforts operate at the sequence level, both the increasing number of protein structures with unknown functions produced by structural genomics efforts and the steady improvement of structure prediction algorithms has created an opportunity for the use of methods that operate at the level of tertiary structure. We present an extension to the FEATURE algorithm (Wei and Altman 1998) that uses multiple local models to characterize an active site or ligand binding site. FEATURE models a site of interest by identifying physicochemical attributes in a series of concentric shells around the active site that are over- or underrepresented with respect to the background. This process produces a model that can be use calculated the likelihood that a prospective environment in a protein structure performs the biochemical function represented by the model. Because the physicochemical attributes are averaged in each spherical shell, orientation of the training sites is unnecessary and variation in the placement of atoms within each shell is permitted.

For larger ligands, a single spherical environment may not be sufficient to capture all of the features relevant to binding and substrate specificity. Hence, we have developed an automated algorithm to select multiple centers on the ligand at which FEATURE models are built. The selected centers are those which best distinguish the ligand binding site from decoy sites at similar atom densities. In order to increase the specificity of the search, one may require a high scoring match to each model, while the sensitivity may instead be increased by allowing a strong match to one model to compensate for a weak match to another.

2. RESULTS

We evaluate this method by applying it to ATP binding sites and show that the combination of two models centered at the C1 carbon of the ribose and the first phosphate atom outperforms either individual model. Furthermore, our modular approach allows us to recognize ATP binding sites even when conformational changes have taken place, as is often the case in the absence of ATP. Figures one and two show histograms of scores obtained for the C1 and PA models, respectively, when scanned against twenty ATP binding proteins whose structures were solved in the absence of ATP (red bars). The blue bars represent scores for a random sampling of five hundred sites in protein structures with atom densities in the ranges expected for an ATP binding site. When the two models are combined (figure 3), the separation between the scores for the twenty apo proteins and the decoy sites improves.

3. FIGURES

Figure 1

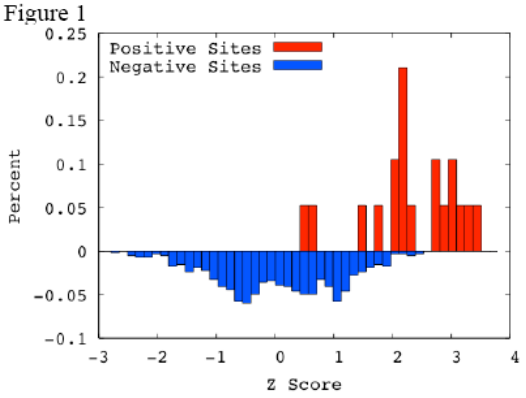


Figure 2

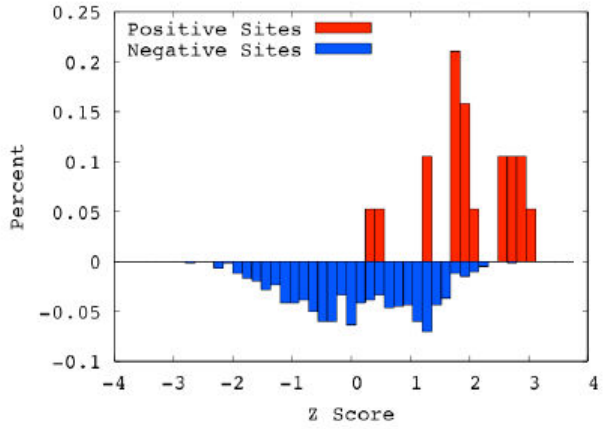
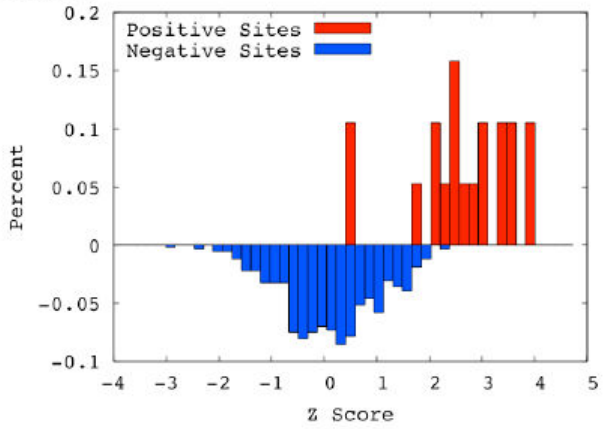


Figure 3



4. REFERENCES

1. Wei L, and Altman RB.1998. Recognizing protein binding sites using statistical descriptions of their 3D environments. Pacific Symposium on Biocomputing. 497-508.

Automated Prediction of Phosphorylation Sites: Existing Methods and the Role of Structure

Shirley Wu [2], Megan Y. So, Russ B. Altman [1]*
1Stanford University Department of Genetics, 300 Pasteur Drive, Mail code: 5120, Stanford, CA 94305
2Stanford Biomedical Informatics, MSOB X-215, 251 Campus Drive, Stanford, CA 94305
*To whom correspondence should be addressed: russ.altman@stanford.edu

1. INTRODUCTION

Phosphorylation in eukaryotes is widespread, and the study of kinases and their corresponding phosphorylation sites has been especially fruitful for drug design. More than 1/3 of the human genome is estimated to encode phosphorylatable proteins, but only a few thousand phosphorylation sites are currently known (4). Discovering phosphorylation sites in vivo is difficult, and accurate prediction algorithms would be an enormous contribution.

A number of computational tools are available, varying in data types and training algorithms used, kinase specificity, and performance. Almost all are based on local sequence information only (1,3,4). We compare these methods and suggest areas for further study based on notable gaps in approaches examined and the following driving biological questions: How do kinases recognize their substrates? How can existing methods for predicting phosphorylation sites be improved? Can structural information improve the quality of prediction? Can information about related sites – such as substrate recruitment sites and docking sites – be integrated to form more accurate and comprehensive predictive models of phosphorylation?

The FEATURE system (5), previously developed by this group, is a supervised machine learning algorithm that characterizes the 3D physicochemical environment around sites of interest, and will be used to develop structure-based models of phosphorylation sites. Preliminary models and studies of 3D atom composition show greater organization in holoproteins compared to apoproteins, and suggest previously unconsidered structural features, such as flexible, positively charged side chains, as potentially informative for a phosphorylation site model. We propose to continue these in-depth, systematic studies of the structural features both local and distal to phosphorylation sites in order to reveal structural signals that could provide greater understanding into the mechanism of phosphorylation site recognition, and help improve accuracy of a site recognition tool. We will also create compound models of phosphorylation sites by incorporating substrate recognition sites and docking sites, an approach that has not yet been applied in this area. Training and test sets will be built from the Phospho.ELM database of validated sites (2), validated Swiss-Prot annotations, and manually inspected PDB entries.

2. REFERENCES

1. Blom N, et al. 2004. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 4: 1633-1649.
2. Diella F, et al. 2004. Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* 5(79).
3. Iakoucheva LM, et al. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Research* 32(3): 1037-1049.
4. Obenauer JC, Cantley LC, Yaffe MB. 2001. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nature Biotechnology* 19:348-353.
5. Wei L, Altman RB. 2003. Recognizing complex, asymmetric functional sites in protein structures using a Bayesian scoring function. *Journal of Bioinformatics and Computational Biology* 1(1): 119-138. *(Full references list available at poster)*

Analysis of Motifs in the Sialytransferase Protein Family

Arun K. Datta[1, 2, 3*], [1] National University Community Research Institute (NUCRI), [2] School of Engineering and Technology (SOET), National University, 4121 Camino del Rio South, San Diego, CA 92108;
[2] UC Humanities Research Institute, University of California, Irvine, CA;
*To whom correspondence should be addressed: adatta@nu.edu

1. INTRODUCTION

Neuraminic acid (Neu5Ac) and its homologs, commonly known as Sialic acids ([1]) are often found as a constituent of the extracellular glycoconjugates ([2]). These 9-carbon negatively charged carboxylated sugars are increasingly recognized as the key determinants of a diverse oligosaccharide structures involved in a large variety of biological events as diverse as cell-cell interaction to oncogenic transformation ([3],[4]). The transfer of sialic acid to such diverse carbohydrate structures is mediated by sialyltransferases (ST), a group of enzymes that transfers sialic acid from its common activated nucleotide sugar donor, CMP-NeuAc. Each of these enzymes is specific for the synthesis of a single linkage and exhibits remarkable specificity for the acceptor substrates. By transferring sialic acid with such strict specificity for both the underlying oligosaccharide structures and anomeric configuration of the substrates, these enzymes generate oligosaccharide structures with a wide diversity. Sialyltransferases of vertebrate origin are all type II membrane proteins and are found primarily in the Golgi apparatus of cells involved in post-transnational sialylation of proteins while in transit through Golgi from Endoplasmic Reticulum ([5]). By now, there are about 720 entries in the CAZY database for these enzymes (family 29; see: http://afmb.cnrs-mrs.fr/CAZY/GT_29.html for updated information) from various sources. These entries account for Neu5Acα2,6Gal (ST6Gal I & II), Neu5Acα2,3Gal (ST3Gal I – VI), Neu5Acα2,6GalNAc (ST6GalNAc I – VI), and Neu5Acα2,8Neu5Ac (ST8Sia I – VI) (for nomenclature, see [6]) often with multiple entries for a single type. So far, total 20 members of this enzyme family with distinct carbohydrate linkage specificity have now been cloned. All of these enzymes have a common structural feature: these are all type II transmembrane glycoproteins with a short N-terminal cytoplasmic domain, a hydrophobic signal-anchor sequence that serves for the membrane spanning, a proteolytically susceptible "stem" region that separates the long luminal catalytic domain from the transmembrane region (Figure I). Comparative peptide sequence analysis of these cloned mammalian enzymes showed the presence of four conserve sialylmotifs in the catalytic domain, namely 'L-'(for long), 'S-'(for short), '-III' (for being third position in sequence) and '-VS' (for very small), which are common to all of this protein family. Experiments by site-directed mutagenesis ([7]) showed the evidence that these motifs contribute to the binding of either donor or the acceptor or both ([8]). While the L-sialylmotif contributes to the binding of the donor substrate ([9]), the motifs -III and –VS contribute to the binding of the acceptor substrate ([10]). S-sialymotif, on the other hand, contributes to the binding of both the donor and acceptor substrates ([11]). Experimental evidence also showed the presence of a disulfide linkage between the L-sialylmotif and the S-sialylmotif ([12],[13]). These two motifs are separated by about 150 amino acids. Apparently this disulfide linkage brings all of these motifs closer together facilitating interaction of these motifs with the substrates. Although there is no structural evidence of any of these mutagenesis studies, a recent report by fold recognition and comparative modeling techniques supports these findings (Figure 2; [14]). Experiments also showed that such proximity of the motifs is essential for maintaining an active conformation of this enzyme family ([12]). In addition, although there is no experimental evidence, comparative sequence analysis also suggests a strong correlation of linkage specificity of these enzymes with the peptide sequence closer to these sialylmotifs ([15], [16]).

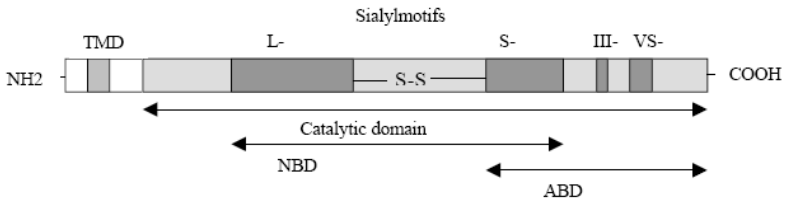


Figure I: Schematic representation of a mammalian sialyltransferase consisting of a short N-terminal cytoplasmic domain, followed by a transmembrane domain (TMD), a stem region of variable length, and the rest is the catalytic domain. The catalytic domain contains four sialylmotifs, namely, L-, S-, -III and –VS. This graphics also represents the position of the nucleotide binding domain (NBD) that spans L-sialylmotif and S-sialylmotif. The acceptor binding domain (ABD) spans sialylmotifs S-, -III and -VS. The presence of a disulfide linkage between the L-sialylmotif and the S-sialylmotif ([12], [13]) is also shown.

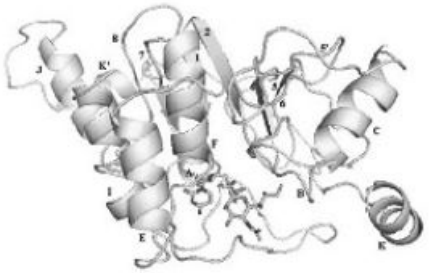


Figure 2. Cartoon diagram of the human ST3Gal I ([14]) modeled using the structure of CstII (PDB id 1RO7). Helices (A, B, C, E, F, I, J, K and K'; in yellow) and strands (β1, β2, β4, β5, β5', β6, β7 and β8; in cyan) have been given the same names as those of corresponding helices and strands in CstII structure. The location of CMP-3-fluoro-NeuNAc (stick representation; carbon, green; oxygen, red; nitrogen, blue) has been derived by superposition of the modeled structure on that of the CstII- CMP-3-fluoro-NeuNAc complex. The structure was rendered using PyMol and reproduced from BMC Struct Biol. 2006; 6: 9.

2. REFERENCES

1. Blix, F.G., A. Gottschalk, and E. Klenk, Proposed nomenclature in the field of neuraminic and sialic acids. Nature, 1957. 179(4569): p. 1088.
2. Varki, A., Diversity in the sialic acids. Glycobiology, 1992. 2(1): p. 25-40.
3. Varki, A., Sialic acids as ligands in recognition phenomena. Faseb J, 1997. 11(4): p. 248-55.
4. Hakomori, S., Possible functions of tumor-associated carbohydrate antigens. Curr Opin Immunol, 1991. 3(5): p. 646-53.
5. Paulson, J.C. and K.J. Colley, Glycosyltransferases. Structure, localization, and control of cell type-specific glycosylation. J Biol Chem, 1989. 264(30): p. 17615-8.
6. Tsuji, S., A.K. Datta, and J.C. Paulson, Systematic nomenclature for sialyltransferases. Glycobiology, 1996. 6(7): p. v-vii.
7. Datta, A.K., Efficient amplification using 'megaprimer' by asymmetric polymerase chain reaction. Nucleic Acids Res, 1995. 23(21): p. 4530-1.
8. Datta, A.K. and J.C. Paulson, Sialylmotifs of sialyltransferases. Indian J Biochem Biophys, 1997. 34(1-2): p. 157-65.
9. Datta, A.K. and J.C. Paulson, The sialyltransferase "sialylmotif" participates in binding the donor substrate CMP-NeuAc. J Biol Chem, 1995. 270(4): p. 1497-500.
10. Jeanneau, C., et al., Structure-function analysis of the human sialyltransferase ST3Gal I: role of n-glycosylation *and a novel conserved sialylmotif.* J Biol Chem, 2004. **279**(14): p. 13461-8.
11. Datta, A.K., A. Sinha, and J.C. Paulson, *Mutation of the sialyltransferase S-sialylmotif alters the kinetics of the donor and acceptor substrates.* J Biol Chem, 1998. **273**(16): p. 9608-14.
12. Datta, A.K., R. Chammas, and J.C. Paulson, *Conserved cysteines in the sialyltransferase sialylmotifs form an essential disulfide bond.* J Biol Chem, 2001. **276**(18): p. 15200-7.
13. Angata, K., et al., *Unique disulfide bond structures found in ST8Sia IV polysialyltransferase are required for its activity.* J Biol Chem, 2001. **276**(18): p. 15369-77.
14. Sujatha, M.S. and P.V. Balaji, *Fold recognition and comparative modeling of human alpha2,3-sialyltransferases reveal their sequence and structural similarities to CstII from Campylobacter jejuni.* BMC Struct Biol, 2006. **6**(1): p. 9.
15. Patel, R.Y. and P.V. Balaji, *Identification of linkage-specific sequence motifs in sialyltransferases.* Glycobiology, 2006. 16(2): p. 108-16.
16. Datta, A. K. Comparative sequence analysis in the sialyltransferase protein family: Analysis of motifs. Current Drug Targets, in press.

JAFA: a Protein Function Prediction Meta-Server

Iddo Friedberg*, Tim Harder and Adam Godzik
Burnham Institute for Medical Research, 10901 N. Torrey Pines Rd., La Jolla, CA 92037, USA
*To whom correspondence should be addressed: idoerg@burnham.org

1. INTRODUCTION

The deluge of genomic data is giving biologists their first post-genomic headache: what do all those genes do? Consequently, there has been a growing effort to computationally determine protein function. Several different methods have been implemented, each with its own strengths and weaknesses. However a biologist trying to predict a proteins function has to know and use many different tools to get an overview of possible functions.

The motivation for creating a function prediction meta-server is twofold. First, the simple concentration of many predictions together in a legible and concise fashion can be very helpful in interpreting function. Second, different prediction methods have different strengths. Combining these strengths may produce a better prediction than any single individual method can.

The Joined Assembly of Function Annotations, JAFA[1] automates the process of querying different function prediction programs, and compiling the results in a legible manner. The user enters the query sequences once, JAFA automatically queries all the different programs, collects and merges the results. JAFA offers its users to fine tune the query using all the options that the original methods would offer, but also provides a useful set of defaults for a quick one-shot. Although the results are merged, JAFA does not try to give a single prediction, but tries to show what the different methods reported and to highlight similarities within these. Yet in order to combine, compare and contrast predictions there is a need for a standardized description of function. Natural language is rife with synonyms and ambiguity, making the comparison of predictions difficult. Therefore JAFA uses the Gene Ontology (GO) [2], a hierarchical annotation which differentiates between the proteins function, the biological process of which it is a part and the cellular component where the protein can be found. GO offers a well defined vocabulary to describe protein function and is supported by a growing number of systems. At this time JAFA queries the following independent function prediction systems: GOtcha [3], GOblet [4], GOFigure [5], InterproScan [6] and Phydbac [7]. Beside those a BLAST [8] search is performed.

JAFA also reports its results in an XML format, and as an electronic spreadsheet, for easy incorporation of the results into larger workflows. A working prototype of JAFA is available on line at http://jafa.burnham.org

2. FUTURE WORK

In the future, we aim to add a structure based prediction methods to JAFA. We are also exploring ways of developing a better consensus scoring method that will reflect the capabilities for the various queried programs. Different versions of JAFA are also in development: one is a stand-alone version of JAFA, to be used for more massive data mining efforts. Another is a Grid-enabled version, which will serve high throughput ventures such as genomics and metagenomics.

Function Driven Target Selection for Structural Genomics

Iddo Friedberg* and Adam Godzik
Burnham Institute for Medical Research, 10901 N. Torrey Pines Rd., La Jolla, CA 92037, USA
*To whom correspondence should be addressed: idoerg@burnham.org

1. INTRODUCTION

Structural genomics is a broad initiative of various centers aiming to provide a complete coverage of protein structure space. However it is not feasible to experimentally determine the structures of all proteins even in a single genome. It is generally agreed that the only viable strategy to achieve such coverage is to carefully select specific proteins ("targets"), determine their structure experimentally, and then use comparative modeling techniques to model the rest. However, the details of the selection strategy are a matter of debate. Here we propose that when selecting targets it should be kept in mind that determining structure is not an ends by itself, but rather a means for understanding the atomic-level implementation of the biochemical function of a protein. What if the protein being modeled has a function different from its template? In that case, there is the very real possibility that no knowledge shall be gained regarding the functional mechanism of the modeled protein. We therefore propose that structural genomics refine the structure-driven approach in target selection by adopting some function-based criteria. We propose to target functionally divergent superfamilies within a given fold group because each requires a structural characterization of its functionality. We have developed a classification system for that purpose and used it to propose a list of additional targets within three functionally rich folds: the TIM barrel, immunoglobulin, and flavodoxin-like folds. We show that a function-driven target selection approach in structural genomics is feasible, and that each of the three folds surveyed has only 50–75% of its functional contents structurally characterized. We call upon structural genomics centers to consider this approach and upon computational biologists to further develop function-based targeting methods for structural genomics efforts. The findings of this study can help pave the way towards a better understanding of structure – function relationships.

2. METHODS

Our approach works as follows: first we demonstrate the utility of SCOP[1] superfamily classification as a proxy for functional classification. We do that using Gene Ontology to SCOP mapping, and applying a Gene Ontology based distance measure between superfamilies. Second we perform an all-vs.-all comparison of representatives of all families in the SCOP database using FFAS03 [2], a sensitive profile–profile alignment tool, which was tested extensively in fold recognition but can also be applied generally to detect distant sequence-based similarities. Third, we use the SCOP-vs.-SCOP assignments to calibrate FFAS03-based similarity, including additional alignment features, and to predict functional difference as measured by proteins being in different SCOP superfamilies and thus having different functions. Finally, we compare representatives of all PFAM[3] families to SCOP families, and using the scoring calibrated in the previous step, pick up matches that are predicted to be similar in fold, yet dissimilar in function (different superfamily). Figure 1 illustrates this approach.

3. FIGURES

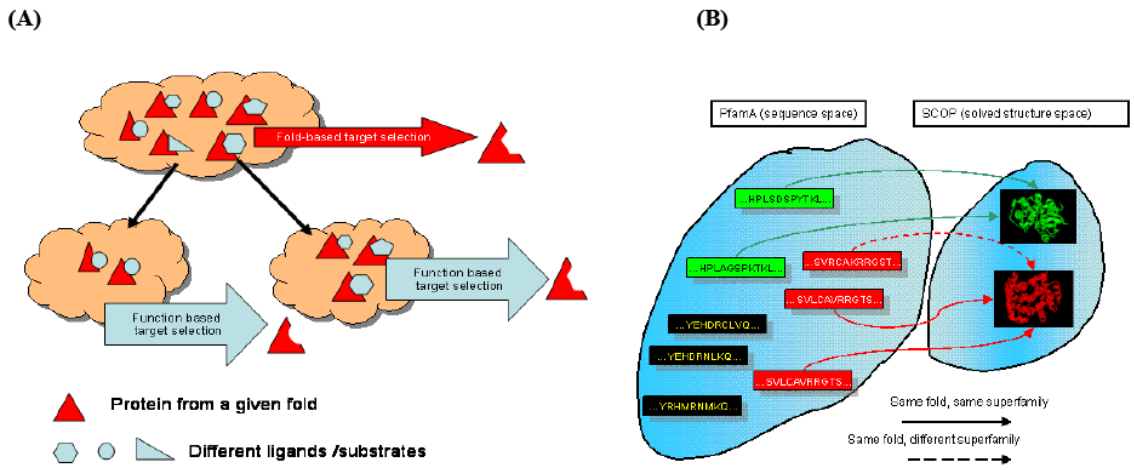


Figure 1: Cartoon illustrating the rationale for function-driven target selection
A. The upper part shows a "classic" target selection strategy for structural genomics: a collection of proteins with the same fold, depicted by a red triangle. A single target for experimental structure determination is picked from this collection. However, the triangle can bind both round and multi-edged objects, which signifies a wide range of functions. The lower part shows the function-based target selection approach: Pick the triangle proteins, but pick one representative from those which bind hexagons and one representative from those which bind circular objects. **B.** Mapping sequence space onto structure space: Representatives of PfamA families are mapped to SCOP folds, using FFAS03. Some can be mapped into a fold but are members of a known superfamily, i.e., their function and structure can be reasonably predicted (continuous lines). Some can be mapped into a fold but cannot be shown to be part of a known superfamily: a known fold but the function is not structurally characterized (dashed line). Others cannot be mapped at all (black background sequences)

4. CONCLUSIONS

In this study we developed a two-way classification that answers the following question: Given that the fold of a protein sequence can be reliably predicted, would this protein have a new function not yet characterized withi that fold? This question immediately begs the following methodological question: how best to know what is a "new" function? We addressed this problem by comparing the different levels of SCOP-based partitioning of proteins to the Gene Ontology annotations of those proteins. The results of this study show recognizing sequences are likely to have a different, uncharacterized function within a known fold is possible, and could be used in a practical selection of targets for structural genomics.

5. REFERENCES
1.Murzin A.G *et al* J. Mol. Biol. (1995) 247, 536-540
2.Jaroszewski L. *et al* Nuc. Acid. Res. (2005) Jul 1;33(Web Server issue):W284-8.
3.Bateman, A. *et al* Nucleic Acids Research(2004) Database Issue 32:D138-D141

The Open Protein Structure Annotation Network

S. Sri Krishna[1,2,3]*, Dana Weekes[1,2], Chris X. Edwards[3], Daniel McMullan[4], Martin Jambon[2], Weizhong Li2, Piotr Kozbial[2], Chloe Zubieta[1,5], Kutbuddin S. Doctor[2], Lukasz Jaroszewski[1,2,3], Adam Godzik[1,2,3] and John Wooley[1,3]

[1] Joint Center for Structural Genomics

[2]Burnham Institute for Medical Research, La Jolla, California, USA 92037

[3]University of California, San Diego, La Jolla, California, USA 92093

[4]Genomics Institute of the Novartis Research Foundation, San Diego, California, USA 92121

[5]Stanford Synchrotron Radiation Laboratory, Stanford University, Menlo Park, California

*To whom correspondence should be addressed: krishna@sdsc.edu

1. INTRODUCTION

Structural Genomics (SG) efforts, in particular, those of the National Institutes of Health (NIH)-sponsored Protein Structure Initiative (PSI), which includes the Joint Center for Structural Genomics (JCSG), have impacted structural biology in many ways, some of which were not completely anticipated. From the time of their inception in the year 2000, the PSI centers have determined high-resolution structures of more than 1200 proteins, with many of them representing novel, previously uncharacterized protein families [1]. In this category, SG centers already outpace the rest of the structural biology field [1,2]. Another major impact of SG efforts came from the rapid progress in developing automated procedures for all steps of protein structure determination, which, increasingly, are being adopted by mainstream structural biology. Both these developments were expected and welcomed by the scientific community. However, the protein architectural information flowing from SG has not been assimilated into mainstream research as rapidly and as widely as that generated by traditional structural biology. We believe the reason for this unanticipated situation is that, unlike traditional structural biology, structure determination at SG centers is inevitably not always - nor even routinely - accompanied by a local stream of connected, synergistic biochemical and biological research. Consequently, the vast majority of protein structures determined by SG centers lack these complementary details and are not described in high impact, peer-reviewed manuscripts, the principal way by which scientists communicate. Instead, the end result of the work of a SG center is usually a set of coordinates deposited in the PDB, information that is not readily assimilated by a typical biologist and opportunities are likely often missed since the scientific application is not recognized. As a result, data from structural genomics is only very slowly absorbed into the wider research stream, largely as correlated experimental data arises.

The goal of our project is to develop "The Open Protein Structure Annotation Network" (TOPSAN; https://www.topsan.org), a radically novel way to collect, share and distribute information about protein three-dimensional structures, and to advance it towards knowledge about functions and roles of these proteins in their respective organisms. TOPSAN will serve as a portal for the scientific community to learn about protein structures solved by SG centers, and also to contribute their expertise in annotating protein function. The premise of the TOPSAN project is that, no matter how much any individual knows about a particular protein, there are other members of the scientific community who know more about certain aspects of the same protein, and that the collective analyses from experts will be far more informative than any local group, let alone individual, could contribute. We believe that, if the members of the biological community are given the opportunity, authorship incentives, and an easy way to contribute their knowledge to the structure annotation, they would do so. Therefore, borrowing elements from successful, distributed, collaborative projects, such as Wikipedia (the free encyclopedia anyone can edit) [3] and from other open source software development projects, TOPSAN will be a broad, collaborative effort to annotate protein structures, initially, those determined at the JCSG. We believe that the annotation of proteins solved by structural genomics consortia offers a unique opportunity to challenge the extant paradigm of how biological data is collected and distributed, and to connect structural genomics and structural biology to the entire biological research community. TOPSAN is designed to be scalable, modular and extensible.

Furthermore, it is intended to be immediately useful in a simplistic way and will accommodate incremental improvements to functionality as usage becomes more sophisticated. Our annotation pages will offer the end user a combination of automatically generated as well as expert-curated annotations of protein structures. We will use available technology to increase the speed and granularity of the exchange of scientific ideas, and use incentive mechanisms that will encourage collaborative participation [4].

Each of the individual PSI centers currently publishes brief, peer-reviewed scientific documents that describe and elucidate (as far as possible) the biologically relevant features of protein three-dimensional structures (Structure Notes) for a small fraction of the structures they solve, while others remain noted only as "uncharacterized hypothetical proteins, to be published" in the PDB. Our goal is to reach out to the general biological community to participate in structure/function annotations of these proteins, with volunteers from the community providing expertise, oversight, validation and management of annotations. The resulting structural biology knowledge repository would be a radical experiment in new ways of collecting, sharing and distributing research information, and will explore ways to modify the traditional, rigid and controlled structure of a research project to accommodate challenges and possibilities brought about by the new, technology-driven, high-throughput science and web-based computer technologies. Rapid advancements in science during the last century have been possible primarily due to the timely communication and sharing of scientific results [5]. Traditional methods of disseminating scientific knowledge, i.e., the publication of manuscripts in peer-reviewed scientific journals and scientific conferences, impede the rate at which scientific information generated at high-throughput centers can be shared and exchanged. In recent years, the internet has proven to be a valuable medium by which information can be exchanged at a rapid pace, and we believe that the TOPSAN project will significantly influence the process of scientific communication.

2. REFERENCES

[1] Chandonia JM, and Brenner SE. (2006). The impact of structural genomics: expectations and outcomes. *Science 311*:347-351.

[2] Sadreyev RI and Grishin NV. (2006). Exploring dynamics of protein structure determination and homology-based prediction to estimate the number of superfamilies and folds. BMC Struct Biol. :6:6.

[3] Wikipedia – The free encyclopedia anyone can edit (http://en.wikipedia.org/wiki/Wikipedia)

[4] Robert Axelrod (1985). The Evolution of Cooperation. Basic Books

[5] Elizabeth L. Eisenstein (1980). The Printing Press as an Agent of Change. Cambridge University Press