Automated Function Prediction SIG Program

Time	Speaker		Title	Page
8:30-8:45			Opening Remarks	
Session 1: Predicting function from structure **Chair: Patricia Babbitt**				
08:45-9:15	Plenary talk: Michael J. Sternberg	Imperial College, London	Prediction of Protein Function from Structure	
09:15-09:30	Nir Ben-Tal	Tel Aviv University	On Protein Prediction using ConSurf, ConSeq and QuasiMotiFinder Web-Servers	
09:30-9:45	Gabriele Ausiello	University of Rome, Tor Vergata	High-Throughput Exploration of Functional Residues in Protein Structures	
09:45-10:00	Dagmar Ringe	Brandeis University	Structure to Function Prediction Methods in Structural Biology	
10:00-10:30	Coffee			
Session 2: Context-based predictions **Chair: Adam Godzik**				
10:30-11:00	Plenary talk: Olivier Lichtarge	Baylor College of Medicine	TBA	
11:00-11:15	Daniela Wieser	European Bioinformatics Institute	Prediction and Contradiction of Protein Function using Annotation Rules	
11:15-11:30	Daisuke Kihara	Purdue University	The Use of Context-Based Functional Association in Automated Protein Function Prediction Methods	
11:30-11:45	Kai Wang	University of Washington, Seattle	Practical Evaluation of Several Automated Function Prediction Methods	
11:45-12:00	Katherine Verspoor	Los Alamos National Laboratory	POSOLE: Automated Ontological Annotation for Function Prediction	
12:00-1:00	Lunch			

Time	Speaker		Title	Page
Session 3: Analysis of functional sites **Chair: Iddo Friedberg**				
1:00-1:30	Plenary Talk: Russ Altman	Stanford University	The FEATURE System for Protein Structure Annotation	
1:30-1:45	Deepak Bandyopadhyay	University of North Carolina at Chapel Hill	Structure-Based Function Inference Using Family-Specific Subgraph Fingerprints Mined from Protein Families	
1:45-2:00	Martin Jambon	The Burnham Institute	SuMo: Structure Comparison of Proteins Focused on Functional Properties of Ligand Binding Sites	
2:00-3:00	Coffee + Poster Session			
Session 4: Functional diversity and specificity **Chair: Russ Altman**				
3:00-3:30	Plenary Talk: Patricia Babbitt	University of California, San Francisco	Pitfalls for Assigning Function from Structural Similarities: Mechanistically Diverse Enzyme Superfamilies	
3:30-3:45	Antonio Del Sol	Fujirebio Inc., Japan	Conformational Changes Playing an Important Role in Promiscuous Protein Functions	
3:45-4:00	Michal Linial	Washington University / The Hebrew University of Jerusalem	Functional Annotations for the Experimental Biologist	
Session 5: Challenging servers				
4:00-6:00	Plenary Talk: Adam Godzik + Assessment Session			

Prediction of protein function from structure

Florencio Pazos, Lawrence A Kelley, Mark Wass & Michael J. E. Sternberg(*)

Structural Bioinformatics Group
Biochemistry Building
Department of Biological Sciences
Imperial College of Science, Technology and Medicine
South Kensington Campus
Exhibition Road
London SW7 2AZ, UK.

*Corresponding author
m.sternberg@imperial.ac.uk

1. INTRODUCTION

Current structural genomics projects are yielding structures for proteins whose functions are unknown. Today there are more than 500 proteins annotated as "hypothetical" in PDB (roughly one per 50 entries). This demonstrates that improvements are urgently required in methods to assign function to proteins both just from their sequence and after the structure has been determined. We report here a new automatic method PHUNCTIONER (1) for structure-based function prediction using automatically extracted functional sequence profiles generate by supervising feature extraction from the Gene Ontology (GO) classification system (2).

Assignment of protein function is complicated , for review see review (3-4). Ultimately one is interested in function at the level of the phenotype but an important step towards this goal is the identification of molecular function. Functional assignment is commonly performed via transfer from the closest homologue of known function and/or via sequence motifs/profiles such as those in INTERPRO. Other methods exploit co-location on the genome, domain co-location, phylogenetic profiles or inferences from the intractome. Recently several groups have developed algorithms to identify functionally important residues often employing sequence conservation and/or structural information including the widely-used strategy known as the evolutionary trace (4). However, evolutionary trace and related approaches primarily focus on the identification of function residues which is distinct from actually assigning a function to the protein.

2. METHODS

The concepts underpinning our approach PHUNCTIONER are:

1. Sequence alignment between proteins with low (<30%) sequence identity are more reliable when based on a structural alignment than from sequence alone

2. GO provides a coherent computational approach to represent protein function at a variety of levels.

3. Go can be used to supervise the grouping of proteins into functional families and this present an alternate approach to supervising classification via a phylogenetic approach.

PHUNCTIONER starts the FFSP (5) structural alignment of a set of protein homologues and uses the GO classification to extract subfamilies with a common GO term. The multiple sequence alignment from a subfamily (derived from the structural alignment) is used to generate a PSSM (position specific scoring matrix). The structure of the protein whose function needs to be assigned (X) would be added to the FSSP multiple alignment and hence the sequence equivalences generated. The sequence of X would then be scored against each PSSM for the alignment and the highest scoring match taken as the prediction of function.

3. RESULTS AND DISCUSSION

To benchmark the approach, we applied a leave-one-out strategy on the FSSP database. We obtained 4,753 sub-alignments (profiles) comprising 121 different GO terms in different levels of the GO. For comparison we contrasted the accuracy of function assignment against that from inheritance from the closest homologue (SEQID) within the FSSP multiple alignment. The accuracy of the FUNCTIONER method ranges from 75% to more than 90% depending on the parameters whereas the one of SEQID goes from 60 to 90%. The results show that PSSM can assign reliably function in zones of low sequence identity where SEQID fails. A "sign test" demonstrates that PHUNCTIONER is significantly better than SEQID below 20% sequence identity (data not shown). As expected, the accuracy of SEQID improves as we permit hits with more similar sequences (from 15% to 30%). For the 30% sequence identity cutoff, the accuracy of both methods is comparable.

In the generation of the PSSM, PHUNCTIONAR identifies the residues that are responsible for a specific function and we are planning to explore the use of these residues to create libraries of 3D templates for function identification on the lines of work by others (e.g. 3).

4. REFERENCES

1. Pazos, F., and Sternberg, M.J.E. 2004. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A* **101:** 14754-14759

2. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32 Database issue:** D258-261.

3. Laskowski, R.A., Watson, J.D., and Thornton, J.M. 2003. From protein structure to biochemical function? *J Struct Funct Genomics* **4:** 167-177.

4. Lichtarge, O., Yao, H., Kristensen, D.M., Madabushi, S., and Mihalek, I. 2003. Accurate and scalable identification of functional sites by evolutionary tracing. *J Struct Funct Genomics* **4:** 159-166.

5. Holm, L., and Sander, C. 1996. The fssp database - fold classification based on structural alignment of proteins. *Nuc. Ac. Res.* **24:** 206-209.

The FEATURE System for Protein Structure Annotation

Russ Altman*, Michael Liang, Shirley Wu
Stanford University
http://feature.stanford.edu/

*To whom correspondence should be addressed: russ.altman@stanford.edu

The last decade has seen the emergence of high-throughput methods for generating biological data, including genome sequencing, microarray expression analysis, proteomics analysis and others. These methods produce valuable information, but pose informatics challenges for indexing, annotation and retrieval. Failure to provide supporting informatics capabilities jeopardizes our ability optimally to use these data. The pipeline of biological structure has been a slow one because of the great expense and technical difficulty in determining three-dimensional structures. Inspired by the success of other high-throughput technology development efforts, the structural biology community has embarked upon a program of "structural genomics" for high-throughput determination of 3D structure using X-ray crystallography or NMR. The analysis of the resulting data is a major challenge; as 3D structures are determined in a non-hypothesis directed manner, methods are needed to search for active sites, binding sites, and other functional/structural sites that provide an initial framework for understanding the function of the molecule. This is the primary focus of our work.

In particular, we have developed a robust statistical method, FEATURE, for representing models of sites in macromolecules. We call these models "three-dimensional motifs" (3Dmotifs) because they capture the 3D arrangement of biochemical and biophysical properties that define a site. We have shown that our 3Dmotifs are useful for identifying key three-dimensional features of sites, and for recognizing these sites in unannotated structures.

The key features of FEATURE program include:

- It considers the location of atoms, chemical groups, and other derived biochemical and biophysical properties (in addition to amino acids) in creating a description of a site, thus enabling features to be recognized and represented at different levels of detail.

- It describes sites with radial symmetry, which allows for better accumulation of statistics with small sample sets, and faster scanning. The loss of precision with radial models is remarkably low.

- It is based on a straightforward statistical model, with built-in control "non-sites" that provide the statistical background, so that significant features can be recognized against this background using a Bayesian scoring scheme.

- It can be extended to include new properties, and has already been extended to work with RNA and fibril proteins, as well as globular proteins.

We have used FEATURE to create the following site models manually:

- Calcium binding site model
- ATP binding site model
- Serine protease active site model
- Redoxin active site model
- EF-hand binding site model
- Magnesium binding (in RNA)
- Disulfide bond environment

We have also used FEATURE to build more than 100 site models using PROSITE patterns as a starting

seed.

Finally, we have results showing the feasibility of:

- Scanning the entire PDB to search for unrecognized sites of interest
- Creating a web and batch interface for annotation of structures in high throughput
- Creating a visualization environment for examining and understanding FEATURE's annotations
- Using FEATURE to assess the ability of structure decoys to predict function

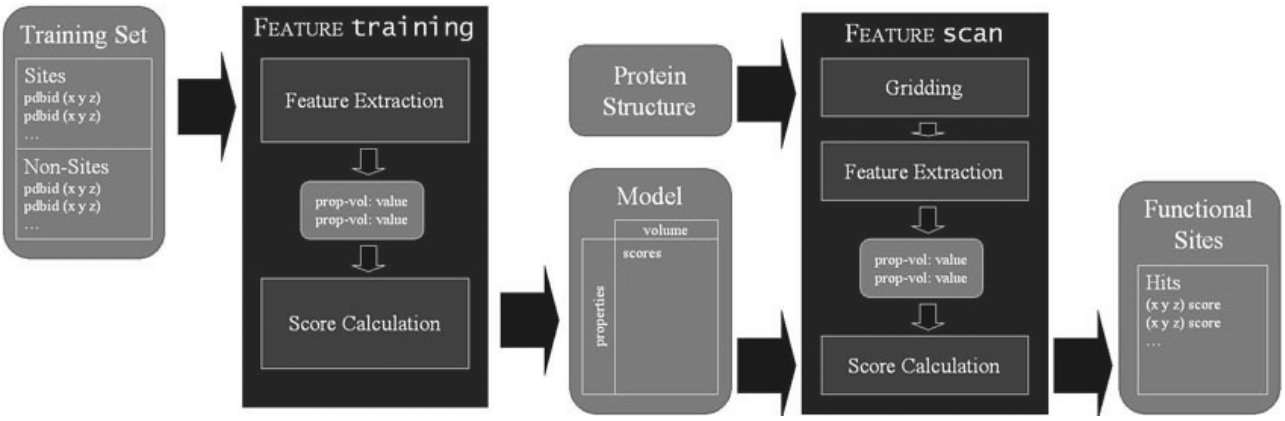


Figure: Overview of FEATURE pipeline. The training step is done once, when volumes and features are defined. The scanning is done each time a structure is annotated with one of the available models.

Table: List of FEATURE's physicochemical properties

AtomName	C ANY	N OTHER	O	S
ChemicalGroup	Hydroxyl RingSystem	Amide Peptide	Amine	Carbonyl
AtomProperties	VDWVolume ChargeWithHis	Charge Hydrophobicity	NegCharge Mobility	PosCharge SolventAccessibility
ResidueName	ALA CYS HIS MET THR HOH	ARG GLN ILE PHE TRP OTHER	ASN GLU LEU PRO TYR	ASP GLY LYS SER VAL
ResidueProperties	Hydrophobic Basic	Charged Acidic	Polar	NonPolar
SecondaryStructure	3Helix Strand Het	4Helix Turn Unknown	5Helix Bend	Bridge Coil

Acknowledgments. The Feature team currently includes Mike Liang, Jessica Ebert, Inbal Halperin, Shirley Wu and Russ Altman. This work is supported by NIH LM05652.

Availability: FEATURE is available via a web interface at http://feature.stanford.edu/

Pitfalls for Assigning Function from Structural Similarities: Mechanistically Diverse Enzyme Superfamilies

Patricia C. Babbitt
University of California
Box 2550, 1700 4th Street
San Francisco, CA 94143-2550 USA
babbitt@cgl.ucsf.edu

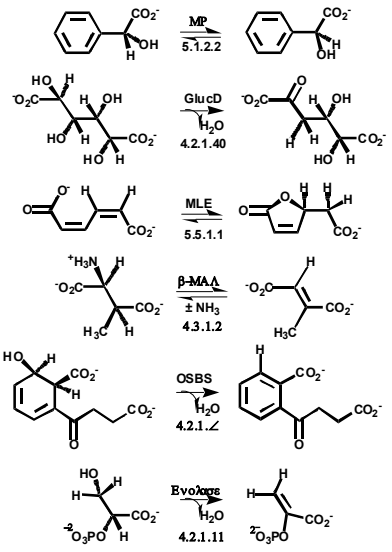
1. INTRODUCTION

The increasing numbers of structures coming available from Structural Genomics and other projects provide enhanced opportunities for using 3D structural comparisons for annotation transfer. In particular, when no statistically significant sequence relationships can be identified, structural comparisons may be required to identify homologs useful for functional inference. But homologs that can only be seen/verified at the structural level may have diverged to mediate very different overall functions. It is difficult to predict the prevalence of this problem because distance metrics used to characterize sequence and structural divergence may not correlate well with functional divergence.

We illustrate these problems for mechanistically diverse enzyme superfamilies in which similarities in overall structure and even in sidechain positions of active site residues may be associated with dissimilar overall reactions which can differ at all 4 digits in the Enzyme Nomenclature (EC) naming hierarchy (1). Analysis of several such superfamilies, each representing many different overall reactions using different substrates and leading to different products, suggests that annotation transfer from structural information for such cases requires understanding the level at which conserved functional characteristics can be correlated with conserved structural characteristics. We find in some diverse enzyme superfamilies that only a partial reaction or chemical capability correlates with similarities in structure at the superfamily level (2, 3). As a result, annotation transfer must be approached with caution for these types of superfamilies. (Figure 1 provides an example.) A solution is proposed in which structure-function correlations are distinguished at the superfamily and family levels, thereby creating a simple hierarchical framework appropriate for annotation transfer in mechanistically diverse enzyme superfamilies (4).

Complications for achieving accurate clustering of families and subgroups within highly diverse superfamilies also contribute to the functional annotation problem. These complications include the observation that different families within a superfamily may have evolved at different rates, that a given function may have evolved more than once within a superfamily, and that connectivity between families can be uneven and therefore difficult to evaluate.

2. FIGURES



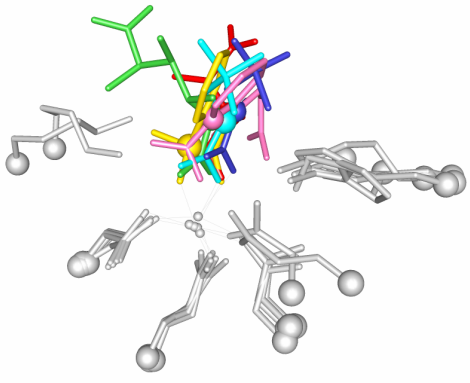


Figure 1. Comparison of active sites and overall chemical reactions performed by members of the enolase superfamily. Left: Some overall chemical reactions performed by divergent members of the enolase superfamily. EC numbers for each reaction are given below the abbreviated name. In each reaction, the conserved partial reaction is abstraction of a proton on a carbon alpha to a carboxylate group in the ligand, leading to a common type of enolate anion intermediate (5, 6). Right: Superposition of active sites showing conserved metal binding ligands and proton abstraction machinery in several liganded structures of divergent members of the enolase superfamily. The sidechains in the similar active sites surround the different ligands of 5 superfamily members whose pairwise sequence identities range from 13-28%. The superpositions represent the MR, GlucD, β-Mal, OSBS, and enolase reactions shown at right. Two liganded structures for enolase are shown, one liganded with substrate and the other liganded with product. No liganded structure for MLE is available.

4. REFERENCES

1. Babbitt, P.C. 2003. Definitions of enzyme function for the structural genomics era. *Current Opinion in Chemical Biology* 7:230-237

2. Babbitt, P.C. and Gerlt, J.A. 1997. Understanding enzyme superfamilies: Chemistry as the fundamental determinant in the evolution of new catalytic activities. *Journal of Biological Chemistry* 272:30591-30594

3. 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

4. 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 (see http://sfld.rbvi.ucsf.edu/)

5. Babbitt, P.C., Hasson, M.S., Wedekind, J.E., Palmer, D.R., Barrett, W.C., Reed, G.H., Rayment, I., Ringe, D., Kenyon, G.L., and Gerlt, J.A. 1996. The enolase superfamily: A general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry* 35:16489-16501

6. Gerlt, J.A., Babbitt, P.C., and Rayment, I. 2005. Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. Archives of Biochemistry and Biophysics 433: 59-70

High-throughput exploration of functional residues in protein structures

Gabriele Ausiello *, Andreas Zanzoni, Daniele Peluso, Allegra Via, Manuela Helmer-Citterich
Centre for Molecular Bioinformatics, Dept. of Biology, University of Rome Tor Vergata,
Via della Ricerca Scientifica, 00133, Rome, ITALY
*To whom correspondence should be addressed: ausiello@cbm.bio.uniroma2.it

1. INTRODUCTION

The detection of local similarities between protein structures may give insights for the identification of a common function. Different tools exist which try to elucidate the mechanisms connecting the similarity of local subsets of residues with the biological activity of whole proteins: i) databases of functionally annotated structures and ii) structural comparison algorithms. Both types of tools suffer from two major problems. The first is the low degree of integration among databases containing functional information. The second is the low or absent integration between existing methods for structural comparison and functional annotation resources.

2. METHODS

We developed a method to integrate many existing databases for 3D functional annotation together with a fast structural comparison algorithm. Ten data sources have been interconnected ranging from solvent exposure to ligand binding ability, location in a protein cavity, secondary structure, functional pattern, protein domain and catalytic activity. All this functional information is bound to the single residue and not to the structure as a whole, permitting to perform detailed queries on the features of single residue sets. All the structural and functional data are stored locally and managed by a fast and powerful database management system that is also able to perform fast and high-throughput local structural comparison. We made this integrated tool available through pdbFun (1), a web server for the structural and functional analysis of proteins at the residue level.

3. RESULTS

PdbFun gives fast access to the whole Protein Data Bank (2) organized as a database of annotated residues. Users can select any residue subset (even including any number of PDB structures) by combining the available functional annotations.
Selections can be used as probe and target in multiple structure comparison searches. For example a search can involve, as a query, all solvent-exposed, hydrophilic residues that are not in alpha-helices and are involved in nucleotide binding. Possible examples of targets are represented by another selection, a single structure or a dataset composed of many structures. The output is a list of aligned structural matches offered in tabular and also graphical format.
This instrument has allowed us to identify cases of convergent evolution in protein structures. Different examples of local structural similarity were highlighted where residues perfectly superposed in 3D are not collinear in the corresponding sequences. In one of these cases the residue order in the sequences is inverted.

4. REFERENCES

1. http://pdbfun.uniroma2.it

2. Deshpande, N., Addess, K.J., Bluhm, W.F., Merino-Ott, J.C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z. et al. 2005. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res*. 33: D233-237

Structure-Based Function Inference Using Family-Specific Subgraph Fingerprints Mined from Protein Families

Deepak Bandyopadhyay[1]*, Jun Huan[1], Jinze Liu[1], Jan Prins[1], Jack Snoeyink[1], Wei Wang[1], Alexander Tropsha[2]
1. Dept. of Computer Science, CB# 3175 Sitterson Hall, University of North Carolina, Chapel Hill, NC 27599, USA
2. School of Pharmacy, CB #7360 Beard Hall, University of North Carolina, Chapel Hill, NC 27599, USA
*To whom correspondence should be addressed: debug@cs.unc.edu

1. INTRODUCTION

We describe a method for predicting the functional family of a protein by *family-specific fingerprints* identified in its structure. These are substructures occurring in most members of a family but rare in a non-redundant subset of PDB, the *background*. Depending on the number of fingerprints found, we assign the structure to a family with a confidence value.

Existing methods for function inference are based on sequence, global structure (fold) similarity, or local structure similarity. *Sequence* methods based on alignment, sequence patterns and phylogeny are most widely applicable, but often not sensitive enough and may miss similarities not conserved in the sequence. This is often a problem for function inference of structural genomics targets, as they are deliberately selected to minimize sequence identity with existing proteins in order to span fold space. *Global structure similarity* is not required for functional similarity (e.g. prokaryotic and eukaryotic serine proteases), nor does it imply functional similarity (e.g. divergent TIM barrel families).

Some *local structure similarity* methods search for known functional sites represented as pockets(3), clefts (7), or surfaces(4); they may fail if there is distortion or mutation in the functional site. Other methods search for known patterns using a graph representation, or discover patterns of limited topology (eg. cliques) from groups of proteins(8). **Our method** assumes no knowledge of functional sites, and can find patterns of arbitrary topology. Each fingerprint is highly specific to its family, and our consensus approach using multiple fingerprints improves the accuracy and specificity of family assignment.

2. ALGORITHM

Our function inference method incorporates the following steps:

1) *Select families* of non-redundant proteins from any classification scheme such as SCOP or EC, or as defined by the user. We processed 31 EC families and 114 families/superfamilies from SCOP version 1.65.

2) *Represent protein structures as graphs*, with nodes at each residue, and contact between residues defined using the sparse *almost-Delaunay*(2) edges, designed to find patterns quickly in data with coordinate perturbations. Distance constraints between non-contacting residues ensure consistent geometry in patterns.

3) *Mine family-specific fingerprints* using the Fast Frequent Subgraph Mining method(5). Fingerprints are defined to occur in $\geq 80\%$ of the family (support), and $\leq 5\%$ of the background (background occurrence).

4) *Search for fingerprints* in a new structure, using subgraph isomorphism sped up by a graph index.

5) *Assign significance* to the function inference, by counting the family fingerprints found in step 4 and assigning a *p*-value based on the distribution of fingerprints in background proteins.

Details of steps 1–3 above may be found in our previous paper(5), and steps 4–5 in a technical report(1).

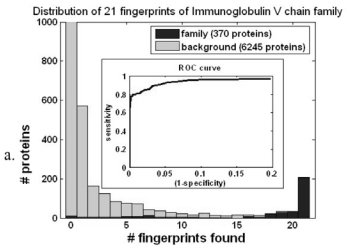
3. EXPERIMENTS AND RESULTS:

We examined family specificity of the fingerprints, validated the method by examining well studied cases of functional similarity, and applied it to function inference of structural orphans.

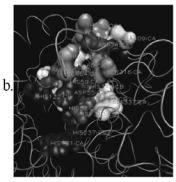
Fingerprint occurrence in background: To assign a *p*-value to a function inference based on how many fingerprints are found in a query protein, we analyze the distribution of fingerprints in the background, as shown in Figure 1(a) for the Immunoglobulin family light/V chain (SCOP: 48727). We use an empirical *p*-value calculation: picking a cutoff point for inferring family membership, we can determine the rate of true and false positives and negatives, calculate specificity (*p*-value) and sensitivity, and draw ROC curves as shown in the inset of Figure 1(a). Often, this process reveals new family members; e.g. 48 proteins having 18 or more fingerprints of Immunoglobulin V chain, do have this domain in the SCOP 1.67 classification.

Validation on SCOP: SCOP families are known to be evolutionarily (and usually functionally) related. Thus, to test the validity of our method, we used the fingerprints from SCOP 1.65 to classify 123 proteins newly added to 41 families in SCOP 1.67. Overall we observed 77% recall when inferring function at 95% specificity points; recall was low for families with few fingerprints, or new members differing in function.

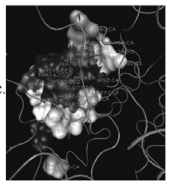
Validation on known cases: For positive validation, we used structural genomics proteins with known or previously inferred function, and cases of functional similarity drawn from the literature. As a negative control, we demonstrated mutual discrimination of TIM barrel families with different functions.



a.



b.



c.

d.

Fingerprints of family	Enriched in family
Multidomain cupredoxin	Cu,Zn and Fe,Mn superoxide dismutases
YjgF-like proteins	Cobalamin (vitamin B12) dependent enzymes
Cation tolerance (CutA)	Mb-cofactor binding
Prokaryotic serine protease	Subtilases
Adenylyltransferases	Class-II AA-tRNA synthetases
Antibiotic resistance	Quinoprotein alcohol dehydrogenases, Lipases

Figure 1 (a) Distribution of Immunoglobulin V chain fingerprints in the background (light bars), and within the family (dark). Inset: ROC curve showing specificity/sensitivity at different cutoff points. (b)-(c) Example of function inference: residues covered by metallo-dependent hydrolase fingerprints in (b) 1m65 (unknown) and (c) 1nfg (family). Color code: white=h'phobic, grey=polar, dark=charged (d) A few functional neighbors without structural similarity found by SCOP enrichment evaluation.

Function inference for structural genomics: Next, we classified Structural Genomics targets in the PDB as proteins with known function, putative function suggested by overall structure similarity, and unknown function with no overall similarity. Our method suggested function assignments for proteins in the last two categories. For example, the function of protein 1npy was inferred from structural similarity to shikimate dehydrogenases (sequence ID 22%) and confirmed by finding 54 of 204 fingerprints, i.e. 98% specificity of function inference. Other cases had no sequence or structural similarity with proteins of known function. For example, the Ycdx protein (PDB 1m65, CASP5 target T0147), which has a rare $(\beta\alpha)_7$ barrel fold. We inferred that this protein has a metallo-dependent hydrolase function, with 30 of 49 fingerprints from SCOP superfamily 51556 of the TIM barrel fold. Our inference was validated by active site template matches and suggestions by the CASP5 target classifiers(6). After fingerprint-based inference, we visualized residues included in fingerprints to verify the *biological significance* of local structures they match; the residues are localized in space and show similar geometry and chemical properties in family and target (Figure 1(b,c)).
Enrichment and functional neighbors: We calculate the *statistical enrichment* of fingerprints in nodes of the SCOP and GO hierarchies, using the hypergeometric distribution with a *p*-value cutoff of 10^{-5}. This helps to determine families that share fingerprints, and thus can be considered *functional neighbors*, i.e. structurally dissimilar families that may have related function. Apart from providing functions to test if the function suggested by fingerprints is false, functional neighbors may help characterize entire families of proteins with unknown function. Figure 1(d) shows a few functional neighbors from our current dataset.
Predicted structures: Our method is robust enough to infer the functional family of predicted structures; thus, we can check if a homology modeled structure infers the template structure's family, to decide if a different template should be chosen. We found that over 25% of top-ranked fold recognition predictions for a typical CASP5 target (eg. T0147) infer the native structure's function, not that of an incorrect template.
Discussion: We present a robust method based on automatic generation of structural fingerprints to suggest a functional family for proteins of unknown function, independent of sequence patterns, structure alignment and templates of known functional sites. Our method may be used for sequence-based function prediction using high quality predicted structures or sequence patterns derived from sequence-ordered fingerprints.

4. REFERENCES

1. Bandyopadhyay, D., Huan, J., Liu, J., Prins, J., Snoeyink, J., Tropsha, A., and Wang, W. 2005. Protein function identification by fast subgraph isomorphism using structure-based fingerprints. *UNC Tech. Report*.
2. Bandyopadhyay, D. and Snoeyink, J. 2004. Almost-Delaunay simplices: Nearest neighbor relations for imprecise points. In *ACM-SIAM Symposium on Discrete Algorithms*, pp. 403–412.
3. Binkowski, T. A., Adamian, L., and Liang, J. 2003. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J Mol Biol*, 332:505–526.
4. Ferre, F., Ausiello, G., Zanzoni, A., and Helmer-Citterich, M. 2004. SURFACE: a database of protein surface regions for functional annotation. *Nucl.Acids. Res.*, 32(90001):D240–244.
5. Huan, J., Wang, W., Bandyopadhyay, D., Snoeyink, J., Prins, J., and Tropsha, A. 2004. Mining protein family specific residue packing patterns from protein structure graphs. *Proc. RECOMB 2004*, pp. 308–315.
6. Kinch, L.N., Qi, Y., Hubbard, T.J. and Grishin, N.V. 2003. CASP5 target classification. *Proteins*, 53 Suppl 6:340–351.
7. Laskowski, R.A., Luscombe, N.M., Swindells, M.B., and Thornton, J.M. 1996. Protein clefts in molecular recognition and function. *Protein Sci*, 5(12):2438–2452.
8. Wangikar, P., Tendulkar, A., Ramya, S., Mali, D., and Sarawagi, S. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J Mol Biol*, 326(3):955–78, 2003.

On Protein Function Prediction Using the ConSurf, ConSeq and QuasiMotiFiner Web-Servers

Nir Ben-Tal*,
Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel

*To whom correspondence should be addressed: bental@ashtoret.tau.ac.il

1. INTRODUCTION

Three web-servers that were developed in my lab over the past 4 years will be critically reviewed: ConSurf, ConSeq and QuasiMotiFinder. A short description of these tools is provided below.

ConSurf (http://consurf.tau.ac.il/)
Key amino acid positions that are important for maintaining the 3-dimensional structure of a protein and/or its function(s), e.g., catalytic activity, binding to ligand, DNA or other proteins, are often under strong evolutionary constraints (1). Thus, the biological importance of a residue often correlates with its level of evolutionary conservation within the protein family. ConSurf is a web-based tool that automatically calculates evolutionary conservation scores and maps them on protein structures via a user-friendly interface (2-4). Structurally and functionally important regions in the protein typically appear as patches of evolutionarily conserved residues that are spatially close to each other. Version 3.0 of ConSurf will be presented here (5). This new version includes an empirical Bayesian method for scoring conservation, which is more accurate than the maximum-likelihood method that was used in the earlier release. Various additional steps in the calculation can now be controlled by a number of advanced options, thus further improving the accuracy of the calculation. Moreover, ConSurf version 3.0 also includes a measure of confidence for the inferred amino acid conservation scores.

ConSeq (http://conseq.bioinfo.tau.ac.il/)
ConSeq is a web server for the identification of biologically important residues in protein sequences (6). Functionally important residues, which take part for example in ligand binding and protein-protein interactions, are often evolutionarily conserved and are most likely to be solvent-accessible, whereas conserved residues within the protein core most probably have an important structural role in maintaining the protein's fold. Thus, estimated evolutionary rates (3), as well as relative solvent accessibility predictions (7), are assigned to each amino acid in the sequence; both are subsequently used to indicate residues that have potential structural or functional importance.

QuasiMotiFinder (http://quasimotifinder.tau.ac.il/)
Sequence signature databases such as PROSITE (8), which include amino acid segments that are indicative of a protein's function, are useful for protein annotation. Lamentably, the annotation is not always accurate. A signature may be falsely detected in a protein that does not carry out the associated function (false positive prediction; FP), or may be overlooked in a protein that does carry out the function (false negative prediction; FN). A new approach emerged, in which a signature is replaced with a sequence profile, calculated based on multiple sequence alignment (MSA) of homologous proteins that share the same function (e.g., reference 9). Following this approach, which is superior to the simple pattern search, one essentially searches with the sequence of the query protein against an MSA library. An alternative approach will be presented, which is implemented in the QuasiMotiFinder web server (10), which is based on a search with an MSA of homologous query proteins against the original PROSITE signatures. The explicit use of the average evolutionary conservation of the signature in the query proteins significantly reduces the rate of FP prediction compared to the simple pattern search. QuasiMotiFinder also has a reduced rate of FN prediction compared to simple pattern searches, since the traditional search for precise signatures was replaced by a permissive search for signature-like patterns that are physicochemically similar to known signatures. Overall, QuasiMotiFinder and the profile search are comparable to each other in performance. They are also complementary to each other in that signatures that are falsely detected in (or overlooked by) one may be correctly detected by the other.

2. REFERENCES

1. Lichtarge, O. and Sowa, M.E. 2002. Evolutionary predictions of binding surfaces and interactions, *Curr. Opin. Struct. Biol.* 12:21-27.

2. Armon, A., Graur, D. and Ben-Tal, N. 2001. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol*, 307:447-463.

3. 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:S71-77.

4. Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E. and Ben-Tal, N. 2003. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, 19:163-164.

5. Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T. and Ben-Tal, N. 2005. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acid Res*. (in press).

6. Berezin, C., Glaser, F., Rosenberg, Y., Paz, I., Pupko, T., Fariselli, P., Casadio, R. and Ben-Tal, N. 2004. ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* 20:1322-1324.

7. Fariselli, P and Casadio, R. 2001. RCNPRED: prediction of the residue co-ordination numbers in proteins. *Bioinformatics* 17:202-204.

8. Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P. and Bairoch, A. 2004. Recent improvements to the PROSITE database. *Nucleic Acids Res*, 32 Database issue, D134-7.

9. Huang, J.Y and Brutlag, D.L. 2001. The EMOTIF database. *Nucleic Acids Res*., 29:202-204.

10. Gutman, R., Berezin, C., Wollman, R., Rosenberg, Y. and Ben-Tal, N. 2005. QuasiMotiFinder: protein annotation by search for evolutionarily conserved motif-like patterns. *Nucleic Acid Res*. (in press).

Conformational changes playing an important role in promiscuous protein functions

Antonio del Sol*, Jorge Arvesú, Zhiqun Xie
Bioinformatics Research Project
Research & Development Division
Fujirebio Inc.
51, Komiya-cho, Hachioji-shi, Tokyo 192-0031, Japan

*To whom correspondence should be addressed: ao-mesa@fujirebio.co.jp

1. INTRODUCTION

The term 'functional promiscuity' has frequently been used to describe the appearance of a new protein function, without loss of the original function, which has been evolutionarily maintained. Promiscuity is not a new concept but the principles governing its mechanism are not fully understood. Indeed, several types of functional promiscuity have been characterized (cross-reactivity, moonlighting, multi-specificity, poly-reactivity) (1), and different protein properties have been suggested that facilitate the ability of a protein to exhibit more than one function. For example, a significant number of aromatic residues in the binding site might be an important factor for cross-reactivity using hydrophobic stickiness, or, highly flexible regions or the occurrence of unrestrictive residues such as glycine, might induce conformational changes leading to the binding of unrelated ligands (2). In general, promiscuity is assumed to involve primarily hydrophobic and other entropy-driven interactions, however, to make things more complex, recent studies show that a single antibody (SPE7) binds specifically several cross-reactants by forming specific hydrogen bonds with each of them (2).

Different authors have addressed the relationship between conformational changes with the appearance of a promiscuous activity (2,3). Despite this fact, due to the complexity of this problem no general methodology dealing with the aforementioned relationship has been proposed. Here we carried out a systematic analysis of several examples of proteins with known experimental data on promiscuous activities (*N*-acetylneuraminate lyase, Myoblobin, Phosphotriesterase, TEM-1 β-lactamase, Aspartate aminotransferase, Taq polymerase I, 3-keto-L-gulonate 6-phosphate decarboxylase, Cytochrome c, Cytochrome P450, Subtilisin, Carbonic anhydrase II, Human estrogen receptor α). Furthermore, we introduced a method based on the representation of protein structures as residue interacting networks in order to characterize those amino acid residues playing an important role for the promiscuous activity.

Previous studies have stressed the fact that independently of the location and type of the mutation in different protein structures the regions that move tend to be generally the same (3). Based on this observation, we considered for each example a multiple structural alignment involving the wild type protein and a set of conformers (small number of mutations, different ligands, different crystals) in order to determine the correlated motions between residues in the structures. Using the residue RMSD correlation coefficient (displacement correlation coefficient) we showed a clear correlation between regions including the promiscuous mutations and different active site residues participating in the promiscuous function. This fact suggests –among other possibilities- that these mutations can be important for keeping the necessary plasticity of the active site as well as for the stability of the protein structure (Fig. 1).

Although the previous analysis gives some insight into the conformational changes associated with the promiscuous function, the residue RMSD essentially expresses the conformational change of each residue, but it can obscure the fine structural details important for the promiscuous activity. A more refined study involved a network representation of protein structures and the analysis of the change of certain topological characteristics between the perturbed and wild type structures. Namely, the number of original residue contacts in the wild type conserved in the perturbed structures was found to be a suitable local topological characteristic describing the sophisticated conformational changes associated with the promiscuous function. Indeed, the amino acids playing an important role for the promiscuous function exhibit a significant change in the previously mentioned topological characteristic (Fig. 2). Further analysis revealed that this change was generally due to the modification in the class (hydrophobic/hydrophilic/glycine) of the residue contacts in the perturbed structures with respect to the wild type structure.

2. FIGURES

Figure 1. Displacement correlation coefficient analysis illustrated with the example of the enzyme TEM-1 β-lactamase. Under three mutations (E104K/M182T/G238S) this enzyme acquires a promiscuous activity consisting in the hydrolysis of a novel type of antibiotics (cefotaxime). Position 182 is located far from the active site and shows a strong displacement correlation with the majority of residues in the active site. This mutation has been reported to be important for increasing the stability of the structure (4).

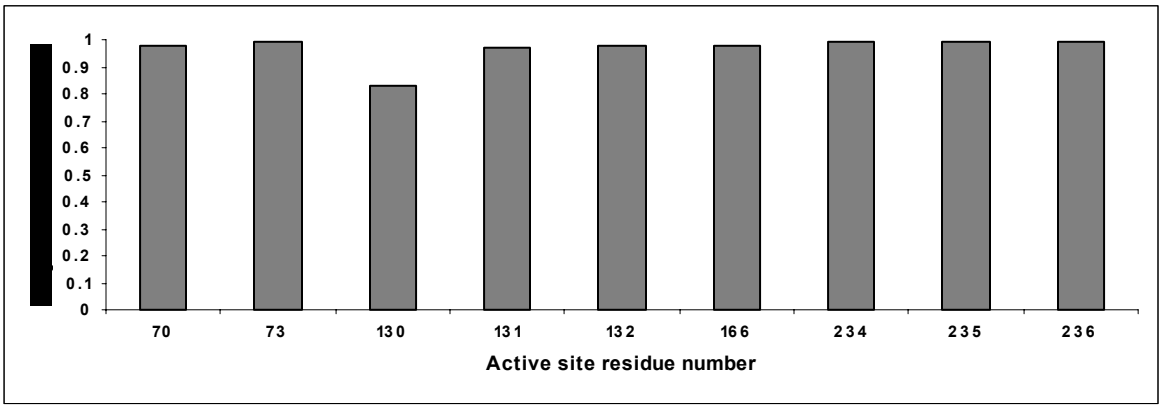
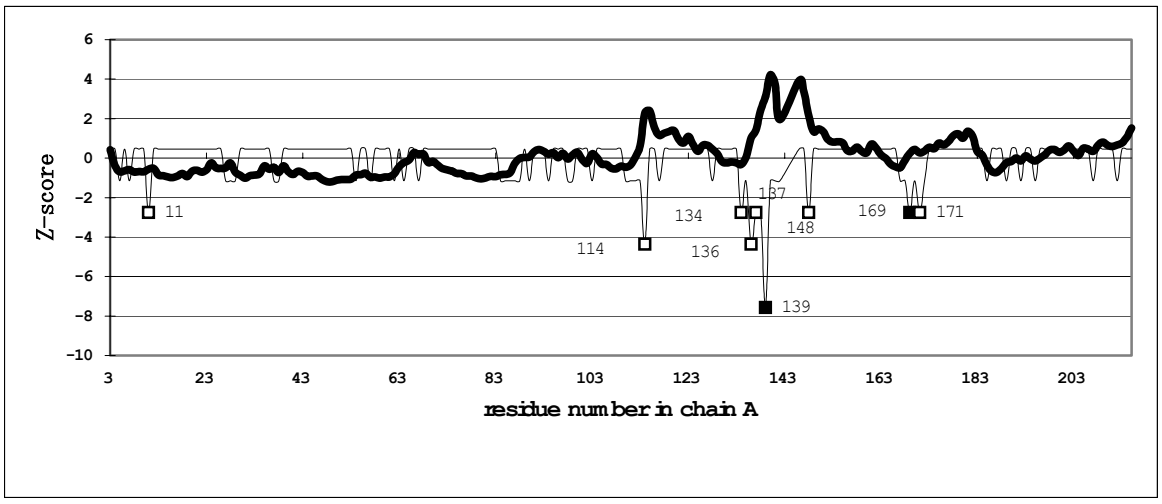


Figure 2. Comparison between the Z-score values of the residue RMSD (thick continuous line) and the number of original residue contacts in the wild type conserved in the perturbed structures (continuous line) for the example of the 3-keto-L-gulonate 6-phosphate decarboxylase (5). Active site residues exhibiting a statistically significant value of the latter characteristic are represented with empty boxes. Active site residues reported in the literature as important for the promiscuous activity are depicted with black boxes (5).



3. REFERENCES

1. James L.C. and Tawfik D.S. 2003. Conformational diversity and protein evolution. *Trends in Biochemical Sciences*, 28:361—368.
2. James L.C. and Tawfik D.S. 2003. The specificity of cross-reactivity: Promiscuous antibody binding involves specific hydrogen bonds rather than nonspecific hydrophobic stickiness. *Protein Sciences*, 12:2183—2193.
3. Sinha N. and Nussinov R. 2001. Point mutations and sequence variability in proteins: Redistributions of preexisting populations. *Proc Natl Acad Sci U S A*. 98:3139-3144.
4. Orencia M.C., Yoon J.S., Ness J.E., Stemmer W.P., Stevens R.C. (2001). Predicting the emergen-ce of antibiotic resistance by directed evolution and structural analysis. *Nat Struct Biol*. 8:238-242.
5. Wise E.L., Yew W.S., Akana J., Gerlt J.A., and Rayment I. 2005. Evolution of enzymatic activities in the Orotidine 5'-Monophosphate Decarboxylase Suprafamily. *Biochemistry* 44:1816-1823.

SuMo: structure comparison of proteins focused on functional properties of ligand binding sites

Martin Jambon*
The Burnham Institute, 10901 North Torrey Pines Road, La Jolla CA 92037, USA
*To whom correspondence should be addressed: mjambon@burnham.org

1. INTRODUCTION

Automated tools for the analysis of protein structure make sense to biologists when they answer practical questions. Today's question is "I have the 3D structure of my protein but I am still unable to guess how it achieves some biological activity: what can I do?". Since this question is pretty vague, we will be more specifically concerned by which classes of small molecules might bind the protein, at which site and how.

Two relatively natural strategies can be thought of. The first approach would consist in docking small molecules or fragments of molecules on the surface of the protein and use some rules to determine which interactions are realistic. The second approach, the one which we have chosen here, consist in comparing the given protein structure with experimentally known protein-ligand complexes.

The main difficulty of such an approach is to derive a data model for protein structures which takes into account most of the "well-known" features of protein-ligand interactions. Small ligands (up to 40 non-hydrogen atoms) are usually flexible. Hence some protein sites may bind the same ligand without being superposable but still using the same kind of interactions with the ligand; this poses a limit to the sensitivity of methods which are based on rigid superpositions (1, 2, 4, 5). Unbound hydrogen bond donors, hydrogen bond acceptors, and aromatic rings can be seen as molecular anchors that have a certain geometry. Another important concept is shape complementarity between the protein and the ligand, and this should be taken into account too. SuMo provides practical solutions to all these issues in one unified system (3).

2. HOW SUMO PROCEEDS

In SuMo, protein structures are represented by a network of microsites. Microsites are triplets of so-called chemical groups (fig. 1). Besides practical concerns, a justification for these triplets is that they may bind a ligand with no degree of freedom and achieve stereospecificity. A microsite has a limited size and can be considered as the smallest element responsible for ligand binding. This is at the level of the microsite that the main information is contained. It includes the shape of the protein in and around the chemical groups.

The actual representation of a protein structure is a graph connecting the microsites that have two chemical groups in common (or close to this). The connections between microsites define potentially larger sites, and these connections are used in the core comparison algorithm. This algorithm consists in matching pairs of triplets that are similar according to various criteria, and then group the pairs of matched microsites that are doubly connected consistently so that matched sites of maximum size could be extracted. There is no further filtering, all the comparisons being performed at the level of the microsites and the connections between them.

Predictions are performed by comparing a given protein structure against the database of ligand binding sites from the PDB. The final results consist in lists of matched chemical groups and their 3D representation.

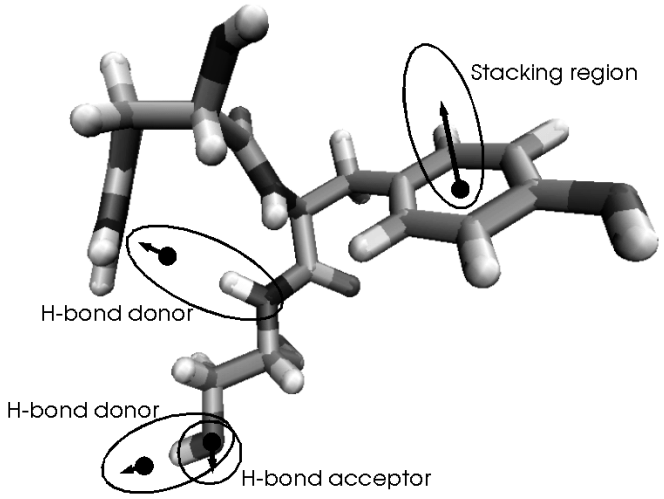


Figure 1: Artistic representation of 3 kinds of chemical groups as defined for SuMo. Black dots with arrows indicate the position of these chemical groups and their geometry. H-bond donors, H-bond acceptors and stacking donors are taken into account only if unbound within the protein. Other chemical groups, not shown here, exhibit different symmetries.

3. IMPLEMENTATION AND AVAILABILITY

SuMo is a fairly complex combination of explicit heuristics, that all have been implemented in Objective Caml, a well-developed programming language (http://caml.inria.fr). Objective Caml combines the expressiveness provided by type inference, garbage collection and full support for functional, imperative and object oriented styles, the safety of static typing, and the efficiency of eager evaluation and a native code compiler. A typical SuMo request consists in scanning the database of more than 20000 ligand binding sites from the PDB. These sites consist of precomputed graphs of microsites which are compared to the query. One such job usually takes about 10-20 minutes on 2 CPUs (2GHz).

SuMo is accessible through its web interface at http://sumo-pbil.ibcp.fr (IBCP, Lyon, France) for non-profit users. It was developed at the IBCP (http://www.ibcp.fr) and at MEDIT (http://www.medit.fr).

4. REFERENCES

1. Barker, J.A. and Thornton, J.M. 2003. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, Sep 1;19(13):1644-9.
2. Ivanisenko, V.A., Pintus, S.S., Grigorovich, D.A. and Kolchanov N.A. 2004. PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucleic Acids Res, 32(Web Server issue)*:W549-54.
3. Jambon, M. 2003. A bioinformatic system for searching functional similarities in 3D structures of proteins. *PhD thesis* : http://martin.jambon.free.fr/phd.html
4. Russell R. B. 1998. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J Mol Biol*, 279(5):1211-27.
5. Shulman-Peleg, A., Nussinov, R. and Wolfson, H.J. 2004. Recognition of functional sites in protein structures. *J Mol Biol*, 339(3):607-33.

The use of context-based functional association in automated protein function prediction methods

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

1. INTRODUCTION

Our thorough analysis of the KEGG Genome collection (1) reveals the extent of the need for efficient and comprehensive methods for computational prediction of protein function. Seventy-five percent of the included genomes contain over 50% gene entries that are ambiguously annotated. Even well characterized organisms (*E. coli*, *C. elegans*, etc.) lack unambiguous annotations in a majority of genes (Fig. 1).

The strongest automated function prediction methods will utilize functional information from all contexts. We have classified current methods into five distinct categories by context: sequence (evolutionary context)- and structure based methods, which utilize global sequence or fold similarity to imply direct homology and extracted sequence or structural motifs and fingerprints to assign distinct functions or localization signals; association (genomic context)-based methods, which use comparative genomics to associate proteins by phylogenetic profiling, co-occurrence, conservation of gene order, domain fusion, expression profiling, and common regulatory elements to imply functional association; interaction (cellular context)-based methods, which assign function or functional association from yeast two-hybrid, co-immunoprecipitation, affinity-purified mass spectrometry experimental datasets; and process (metabolic context)-based methods, which utilize the structured sequential nature of biochemical pathways to assign sequences to yet unidentified reactions.

The current version of our prediction method, implemented in the Protein Function Prediction (PFP) server (http://dragon.bio.purdue.edu/pfp), uses sequence- and structure-based methods. The aim of our sequence-based method is to utilize information from relatively weak hits in PSI-BLAST, which are not conventionally used. Typically, weak hits in PSI-BLAST are not perfect orthologs to the query sequence, but rather share a common functional domain. In addition to simply transferring the function of the common domain to the query sequence, our idea is to also consider those functions which are frequently associated with the annotated functions of the domain. To this end, we have built Function Association Matrices (FAMs) that quantify the co-occurrence of Gene Ontology (GO) annotations in sequences of the UniProt database (Figure 2). The GO is a controlled hierarchical vocabulary describing the function of genes in three categories: function, process, and component (2). Approximately two thirds of associated function pairs mined from UniProt bridge functions of different categories. Thus, we can assign function using FAMs that cannot be retrieved directly from highly similar sequences or structures. Taking advantage of the hierarchical nature of the GO vocabulary, we have developed a series of FAMs in varying "resolution", *i.e.* depth of the functional association in the GO hierarchy. The structure of GO also allows us to define functional proximity as the coordinate distance of the annotation sets of two proteins on the GO tree.

The aim of our structure-based method is to assign annotations from known and predicted structures. We expect that most of the public input into PFP will contain only sequence information rather than both sequence and structure, but structure can still be taken into account through protein structure prediction by threading. Due to the recent improvements in methodology and the rapid growth of databases, we can assign structure to more than 70% of the genes in a complete genome (3). Almost 70% of domain topologies are associated with a unique function; therefore it is likely that we can make an initial function prediction of a gene if it is assigned a single-function topology by threading.

PFP is benchmarked on approximately 400 protein test sequences randomly picked up from UniProt. The benchmark is designed to test how well GO annotations of test proteins are reproduced by PFP and how close the predicted GO terms are to the correct ones.

2. FIGURES

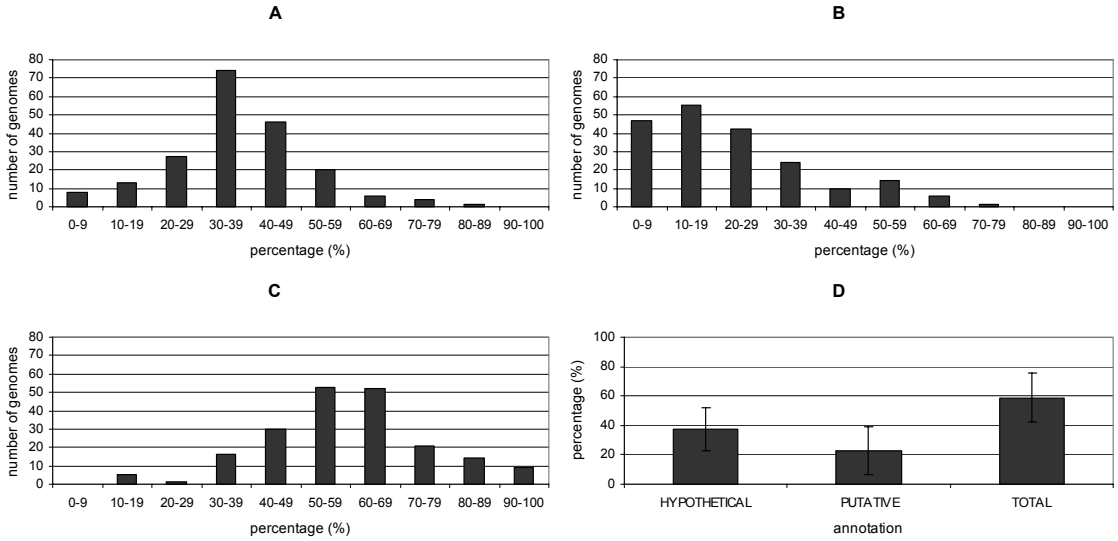


Figure 1. Analysis of the 201 genomes from KEGG: [A] distribution of unknown function annotations among entries (unknown annotations include terms like 'hypothetical', 'unknown', and 'uncharacterized'); [B] distribution of ambiguous annotations among entries (ambiguous annotations include terms like 'putative', 'probable', 'homolog', 'ortholog', 'family', and 'possibly'); [C] distribution of annotations including either unknown or ambiguous terms or both (some entries contain terms that could be applied to both categories); [D] average percentage (among genomes, ± 1 std. dev.) of entries including hypothetical annotations, putative annotations, or both (HYPOTHETICAL: 37.11 ± 14.63%; PUTATIVE: 22.93 ± 16.42%; TOTAL: 58.68 ± 16.77%).

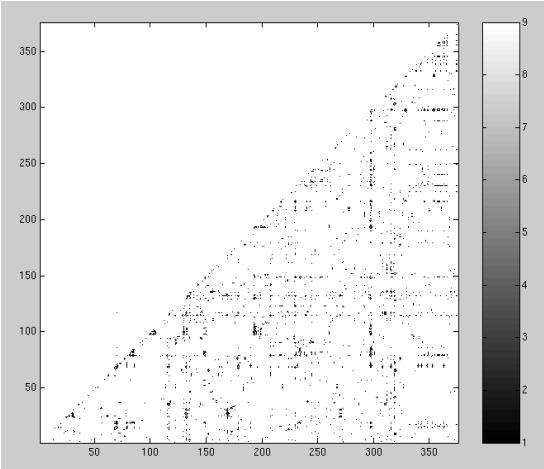


Figure 2. Scaled-down representation of the UniProt FAM: This example shows the straight association score between 375 GO annotations from 1000 UniProt entries at the most specific level (±0 node distance in the GO graph). Darker spots indicate a higher association score.

3. REFERENCES

1. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32 Database issue: D277-D280.
2. Harris, M.A., Clark, J., Ireland, A. *et al*. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32 Database issue: D258-D261.
3. Kihara, D. and Skolnick J. 2004. Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. *Proteins: Structure, Funct. Bioinformatics* 55: 464-473.

Functional Annotations for the Experimental Biologist

Michal Linial[1,2] and Noam Kaplan [1]
[1] Dept of Biological Chemistry, The Hebrew University of Jerusalem, 91904, Israel
[2] Dept of Computer Science and Engineering, University of Washington, Seattle, 98195, WA, USA
michall@cs.washington.edu

1. INTRODUCTION

Numerous high throughput (HTP) methodologies have been developed in recent years. Examples are the two hybrid system, genomic based SNP arrays, DNA microarrays, proteomic based LC MS/MS and protein chips. The application of such technologies drastically changed the scale and the pace of discovery in biological and medical research. The outcome of such large-scale HTP experiments is a subset of proteins or genes that are often associated with quantitative values or some calculated scores. Due to the large scale of these experiments (i.e., the proteome of a human tissue under a defined condition or a comparative gene expression of two tissues), they often result in very large lists of hundreds or thousands of sequences. Naturally, without appropriate tools to infer the functional properties for these sequences, the knowledge from such HTP experiments becomes very limited. Of course, functional annotation provided by experts that is based on manual inspection of thousands of sequences is very slow and is prone to human bias and errors. Herein, we focus on the practical needs of biologists and present directions to accelerate the functional knowledge gain for genes and proteins.

An appreciation for the need for functional annotation initiatives matured with the unprecedented growth in genomic information. Most servers and tools for functional annotations focus on one type of information: genomic sequences, evolutionary or comparative genomics, structural and proteomics or sequence homologies. We propose integrating different annotation sources into one framework, thus avoiding shortcomings of individual sources and increasing biological knowledge in order to enhance the process of experimental result interpretation.

2. RESULTS

An experimental biologist that has accumulated thousands of data records from large-scale experiments faces the following uncertainties: (i) Which reference database is most suitable for the type of data collected? (ii) What is the quality of the functional interpretation? (iii) How should independent sources of information be integrated? (iv) What is the most compact and intuitive way to store, present and retrieve the information? These issues will not be solved by adding more sources of information but instead by presenting statistically sound analysis tools and intuitive visualization solutions.

Several interesting meta-search portals that combine rich sources are effective only when a small numbers of unrelated sequences are to be analyzed. We present here a web server named PANDORA [1] (www.pandora.cs.huji.ac.il) that provides the biologist with an analysis tool combined with an intuitive graphic presentation for a fast and efficient assessment of biological knowledge. The idea behind PANDORA is to provide a highly informative view of a set of proteins and the combinations of annotations on them, while considering data from several different annotation sources. Especially important here is the notion of examining combinations of annotations: While several annotation sources aim at providing standard functional annotation, these sources suffer from significant incoherence, mostly due to methodological differences. By inspecting the overlaps of annotations on proteins, PANDORA makes it easy to overcome this problem. Additionally, several functional groups are characterized by possessing a certain combination of annotations. PANDORA obtains annotations from SwissProt, Taxonomy, InterPro and the hierarchical classification terms from ENZYME, SCOP and GO databases. The annotations that are obtained are treated as binary properties on each of the proteins. Next, PANDORA considers the annotations on a given set of proteins, and constructs a hierarchical concept graph in which the nodes represent subsets of proteins that share a unique combination of annotations and the edges represent intersection and inclusion relations between these subsets. For further information see Figure 1.

Two additional developments in PANDORA increase its usability and generality: (i) An input of quantitative measurements and (ii) an output of the statistical significance for annotations associated with any subset of the proteins.

Most large scale experiments (i.e., large-scale RNAi knockdown, localization data, gene expression, 2D gels proteomics, two hybrid system) are associated with experimental values that reflect quantitative trends. The most recent version of PANDORA (ver. 3.2) allows associating each entry with one or more quantitative values (i.e. differential gene expression, protein length or even BLAST e-values). This allows easy recognition of subsets of proteins that both share biological annotations and behave quantitatively similar (for example, a subset of proteins that are all involved in the same biological pathway and are also differentially up-regulated).

Statistical evaluation is especially important for methods that deal with functional annotation, which can be exceptionally non-coherent and vague. PANDORA offers a statistical evaluation (p-values) for the appearance of annotations on a set of proteins. This evaluation answers the following question: Given that an annotation A appears in our protein set X times, what is the probability of obtaining X or more hits for the annotation when randomly selecting the same amount of protein from a background database? Furthermore, PANDORA currently supports several background scenarios for the statistical analysis, including Affymetrix microarrays and whole-proteome experiments.

One interesting aspect of PANDORA is the tendency of incorrectly annotated proteins to appear separated in the concept graph. This is due to the natural aggregation of biological properties: proteins that are functionally similar usually several annotations. However, proteins that share only one functional annotation but differ greatly in their other annotations are often found to be incorrectly annotated. In a previous study [2] we have exploited this phenomenon, which is often very easy to spot in PANDORA graphs, and showed that it may be used to detect false positive annotations automatically.

We will discuss and demonstrate how PANDORA enhances the biological understanding of large, non-uniform sets of genes and proteins originating from proteomics, genomics and computational sources. Furthermore, the integration of several different annotation sources simultaneously allows gaining insight into biological relations of structure, function, cellular location, taxonomy, domains and motifs.

3. FIGURES

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Kinase	1	0	0	0	0	0	0	1	1	0
Receptor	0	0	0	0	0	1	0	0	0	1
Development	1	1	1	1	1	1	1	1	1	1
Homeodomain	0	1	0	0	1	0	0	0	0	0
Membrane	0	0	1	0	0	1	0	0	0	1
Enzyme	1	0	1	0	0	0	1	1	1	0
Transcription	0	1	0	0	1	0	0	0	0	0
Nucleus	0	1	0	0	1	0	0	0	0	0

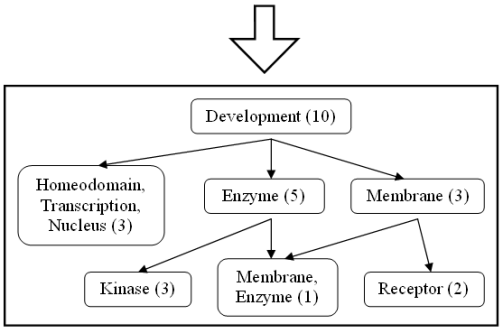


Figure 1. Construction of the concept graph. The annotations on a set of proteins are represented as a binary matrix where the rows are annotations and the columns are the proteins. Each row vector reflects a subset of proteins that share a given annotation. Each of these subsets is a preliminary node in the graph. Next, each of these subsets is compared with the others, and a directed acyclic graph is constructed so that it represents intersection and inclusion relations between them.
The final concept graph possesses three important attributes. First, although individual proteins are not showed in the graph, the annotation matrix can be fully reconstructed from the graph, meaning all the information is present. Second, every subset of proteins that has a unique combination of proteins is represented as a node in the graph. Finally, due to the hierarchical nature of the graph, we obtain the following simple rule: The proteins of a given node possess the annotations that are assigned to that node and to all of its ancestor nodes.

This work is supported by the Sudarsky Center for Computational Biology (SCCB) and by the EU Framework VI DIAMONDS consortium.

3. REFERECES

1. Kaplan N, Vaaknin A, Linial M: PANDORA: keyword-based analysis of protein sets by integration of annotation sources. *Nucleic Acids Res* 2003, 31:5617-5626.
2. Kaplan N, Linial M: Automatic detection of false annotations via binary property clustering. *BMC Bioinformatics* 2005, 6:46.

Structure-to-Function Prediction Methods in Structural Biology

Dagmar Ringe*[a], Mark A. Wilson[a], Ying Wei[b], Mary Jo Ondrechen[b]
[a] Department of Biochemistry, Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA 02454, and [b]Department of Chemistry and Chemical Biology and Institute for Complex Scientific Software, Northeastern University, Boston, MA 02115, USA
*To whom correspondence should be addressed: ringe@brandeis.edu

1. INTRODUCTION

Structural biology in the genomics era faces the challenge of determination of function from 3D structure, the critical next step toward the realization of the promises of genome sequencing. On the order of 10^3 protein structures in the PDB currently are annotated as "hypothetical" or "unknown function" and this number is increasing continuously. The annotation of function is usually dependent on sequence similarity to identify proteins that are expected to be similar in structure and therefore similar in function. Even if the sequence comparison does not find a related protein, the overall structural fold may be similar to one that is already known, but such a structural relationship still does not necessarily identify function. The reason for the discrepancy is that the task of going from structure to function is often far from straightforward. Many proteins with similar and recognizable folds have completely different functions, even sometimes when there is sufficient sequence similarity to consider them 'homologous'. The best examples of this principle are the enzymes having the TIM barrel fold. The types of reactions catalyzed by proteins having this fold are numerous and varied. Alternatively, two proteins may have completely different folds, but catalyze the same reaction, with the same residues and configurations in the active site. A good example is the set of pyridoxal phosphate dependent transaminases of fold types I and II. These proteins catalyze the same reaction in active sites that are practically identical, but structurally they are totally different, having completely different folds.

In addition, the important residues in an active site may not be obvious. Many reactions in biology may be characterized by the steps required to bring about any chemical transformation. The catalytic entities involved in each step, such as acids or bases, can be inferred from the known chemistry. Residues that can play these roles are well defined, however it is not so easy to determine which residues are actually playing these roles. Ideally, a structure with substrate bound would resolve the question, but such structures are rarely available. Therefore, another method is needed to identify catalytic residues and recognition residues. In this paper we present how a computational predictive tool can aid in the identification of the functionally important residues in proteins of unknown function.

The catalytic power of an enzyme relies not only on the nature of the residues that aid catalysis, but also their position relative to the substrate. The method that identifies residues in the active site of a structure therefore also locates their relative positions and defines the type of chemistry that is possible, and potentially the substrate that can be recognized.

2. METHOD

We have previously reported on THEMATICS, a simple and fast computational tool for the prediction of catalytic and recognition sites in proteins that requires only the three-dimensional structure of the query protein as input (1-6). THEMATICS is based on Poisson-Boltzmann calculations of the electrical potential for the protein structure, calculation of the theoretical titration curves (average charge as a function of pH) for all of the ionizable residues, and then statistical analysis of the computed titration curves to identify the ones that deviate the most from typical Henderson-Hasselbalch behavior. Clusters in coordinate space of two or more residues with deviant theoretical titration behavior are considered predictive and indeed predict localized interaction sites in proteins with high recall and high precision. Here we report on how these predictive tools can be used to aid the experimental study of proteins of unknown function. We focus on a family of structurally similar proteins of biomedical importance that apparently have different chemical functions.

3. RESULTS FOR DJ-1 FAMILY PROTEINS

Examples of new structures from the functionally diverse but structurally similar DJ-1 family (7, 8) are

featured in the present paper. Human DJ-1 is a protein of unknown function that apparently plays a neuroprotective role. Mutations of DJ-1 have been associated with Parkinson's disease. YDR533Cp from yeast (of unknown function), the Thij protein from *E. coli* (of unknown function), the hypothetical protein Yhbo from *E. coli*, the chaperone Hsp31 from *E. coli*, and Protease I from *Pyrococcus horikoshii*, are all closely related structures. Human DJ-1 does not exhibit any significant protease activity. For Protease I from *Pyrococcus horikoshii*, THEMATICS predicts a cluster at the protease active site that includes the catalytic triad members C100, H101, and E474. For human DJ-1, THEMATICS finds a different cluster consisting of E15, E16, E18 and D24, located adjacent to, but not overlapping, the corresponding triad site. For yeast YDR533Cp, THEMATICS predicts E30, D57, H108, H139, and E170, a cluster that intersects both the corresponding triad site and the predicted DJ-1 site; H139 and E170 are located in positions corresponding to those of His and Glu of the Protease I triad whereas E30 in the predicted YDR533Cp cluster is structurally aligned with E18 of human DJ-1.

Table 1. THEMATICS predictions for DJ-1 family proteins

Protein/PDB ID	THEMATICS result
Human DJ-1/1SOA	[E15a, E16a, D18a, D24b]
Thij	[E14a, E15a, E17a, D23b, R27b]
Protease I/1G2I	[E12a, D13a, E15a, K43a, E74a, C100a, H101a, D126a, D153a, E212b, D213b, E215b, D353b, E412c, E415c, R471c, E474c, C500c, H501c, D526c]
PfpI/1OI4	[E35a, E38a, H96a, D99a, H126a, H296b, D299b, H326b, D351b]
YDR533Cp/1RW7	[E30a, D57a, H108a, H139a, E170a]
ChaperoneHsp31/1N57	[H74a, E77a, E105a, H155a, H186a, D214a]

Subunits are designated a,b,c. Corresponding positive clusters in the other subunit(s) are implied.

4. CONCLUSIONS

The predicted THEMATICS clusters for the DJ-1 family members enable us to sort them into groups (*e.g.* DJ-1 and Thij) with similar predicted active sites and hence presumably similar function.

The facile identification of binding and recognition sites in proteins with a simple calculation provides important and time-saving clues in the determination of a protein's function.

5. REFERENCES

1. Ko, J., L.F. Murga, P. Andre, H. Yang, M.J. Ondrechen, R.J. Williams, A. Agunwamba, and D.E. Budil 2005. Statistical Criteria for the Identification of Protein Active Sites Using Theoretical Microscopic Titration Curves. *Proteins: Structure Function Bioinformatics* 59:183-95.
2. Murga, L. F., Y. Wei, P. Andre, J.G. Clifton, D. Ringe, and M.J. Ondrechen 2004. Physicochemical methods for prediction of functional information for proteins. *Israel Journal of Chemistry* 44:299-308.
3. Ondrechen, M. J., J.G. Clifton and D. Ringe 2001. THEMATICS: A simple computational predictor of enzyme function from structure. *Proc. Natl. Acad. Sci. (USA)* 98:12473-78.
4. Ringe, D., Y. Wei, K.R. Boino and M.J. Ondrechen 2004. Protein Structure to Function: Insights from Computation. *Cellular Molecular Life Sciences* 61:387-92.
5. Shehadi, I. A., H. Yang and M.J. Ondrechen 2002. Future directions in protein function prediction. *Mol Biol Reports* 29:329-35.
6. Shehadi, I. A., A. Abyzov, A. Uzun, Y. Wei, L.F. Murga, V. Ilyin, and M.J. Ondrechen 2005. Active Site Prediction for Comparative Model Structures with THEMATICS. *Journal of Bioinformatics and Computational Biology* 3:127-43.
7. Canet-Aviles, R. M., M.A. Wilson, D.W. Miller, R. Ahmad, C. McLendon, S. Bandyopadhyay, M.J. Baptista, D. Ringe, G.A. Petsko, and M.R. Cookson 2004. The Parkinson's disease protein DJ-1 is neuroprotective due to cysteine-sulfinic acid-driven mitochondrial localization. *Proc Natl Acad Sci U S A* 101:9103-08.
8. Wilson, M. A., C.V. St. Amour, J.L. Collins, D. Ringe and G.A. Petsko 2004. The 1.8 A resolution crystal structure of YDR533Cp from Saccharomyces cerevisiae: A member of the DJ-1/ThiJ/PfpI superfamily. *Proc Natl Acad Sci U S A* 101:1531-36.

POSOLE: Automated Ontological Annotation for Function Prediction

Karin Verspoor*, Judith Cohn, Susan Mniszewski, Cliff Joslyn
Los Alamos National Laboratory, PO Box 1663, Los Alamos, NM, 87545 USA
*To whom correspondence should be addressed: verspoor@lanl.gov

1. INTRODUCTION

We present the methods utilized in a system aimed at predicting the function of protein targets, as represented by a node in the Gene Ontology (1). The core architecture has been utilized in two distinct contexts: function prediction from text for the BioCreAtIvE evaluation Task 2 (2), and function prediction from protein sequences for the CASP function prediction task (3).

The system we have developed is called POSOLE, or the POSet Ontology Laboratory Environment. POSOLE consists of a set of modules supporting ontology representation, categorization of nodes in the ontology, and analysis. The analysis modules provide support for analysis of the ontological structure, the structure of input queries to the categorization module with respect to that structure, and the structure of the predicted categorization with respect to a given set of expected answers. The system requires the definition of mappers called QueryBuilders for implementation within a specific application. These QueryBuilders define how to map from the relevant input for the application to a set of ontology nodes. For both the BioCreAtIvE and CASP applications, this is done by considering the neighborhood of the protein in the input space and associating entities in the neighborhood to Gene Ontology (GO) nodes. Then POSOLE categorizes the collection of GO nodes based on their distribution in the GO structure, utilizing a technology called POSOC, the POSet Ontology Categorizer (4) (originally called GOC, the Gene Ontology Categorizer (5), but generalized for use with any partially ordered ontology). The resulting set of Gene Ontology nodes is interpreted as the most representative nodes for the function of the input protein. The architecture of the two applications and the common POSOLE modules can be seen in Figure 1.

2. BIOCREATIVE APPLICATION

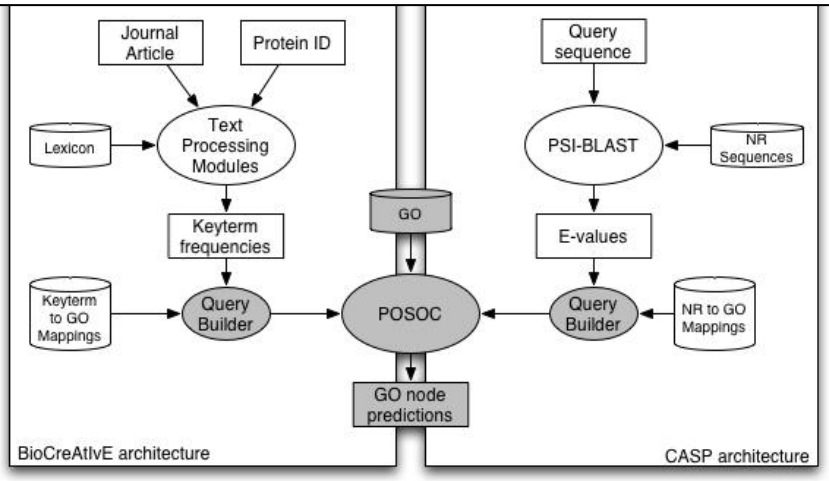
For BioCreAtIvE Task 2, we were provided with a protein identifier (Swiss-Prot identifier) and a relevant journal publication and asked to predict the function of the protein, as represented by a set of GO nodes, on the basis of that publication. The application defines a POSOLE QueryBuilder that is responsible for associating terms in the publication to GO nodes. This is accomplished through the use of natural language processing components. Specifically, the document is processed to morphologically normalize terms to their base forms, identify sentence boundaries, and calculate the relative importance of terms using the statistical measure TFIDF (term frequency inverse document frequency) with respect to a background corpus. We then identify all references to the input protein in the document and collect the terms in a context window around those references. These terms are considered to be in the contextual neighborhood of the protein, and are assumed to be most indicative of the protein's function. These terms in turn are mapped to specific GO nodes through lexical matches between the text and the text of the GO nodes, in the GO node labels and node definitions as well as in additional sets of terms that were previously associated with specific GO nodes via unsupervised learning (see (2)). An input query for POSOC is constructed which consists of the set of matched GO nodes, weighted according to the TFIDF of the matching term.

3. CASP APPLICATION

For the CASP function prediction task, we were provided with a protein sequence and asked to predict the function of the protein, again in terms of a set of GO nodes. The application defines a POSOLE QueryBuilder that is responsible for associating the input sequence to GO nodes. In this case, we use a "nearest neighbor" approach: we identify close neighbors of the input sequence in sequence space and collect the GO nodes associated with those neighbors in a curated data set (Swiss-Prot).

To identify close neighbors of a target sequence, we performed a PSI-BLAST (Position-Specific Iterated BLAST) (6) search on the target against the NCBI NR database, with 5 iterations, using the default e-value threshold of 10. Once the nearest neighbors have been identified, we collect the GO nodes associated with these sequences utilizing the UniProt Swiss-Prot to GO mappings. Finally, we build a weighted collection of GO nodes, where each node in the collection is weighted according to the PSI-BLAST e-value. Several near neighbors of the original target sequence may map to the same nodes. In this case, each occurrence of a GO node will be weighted individually according to its source.

Figure 1: POSOLE application architectures. The architectures for the BioCreAtIvE and CASP protein function prediction applications, built around the core POSOLE modules, shown in grey.



4. POSOC

For each application, the collection of weighted GO nodes becomes the input query to the categorization technology POSOC (4). This technology aims to identify a set of nodes in a partially ordered set, such as the GO, which best summarize or categorize a given list of input nodes. The technology is based on a view of bio-ontologies as combinatorially structured databases, and draws on the discrete mathematics of finite partially ordered sets (*posets*). Briefly (for more detail, see (2),(4)), after identifying the set of input nodes in Gene Ontology space, POSOC traverses the structure of the Gene Ontology, percolating hits upwards, and calculating scores for each GO node. POSOC then returns a rank-ordered list of GO nodes representing cluster heads. In the end, this provides an assessment of which nodes best cover the input set. We consider this set of cluster heads to be indicative of the function of the input protein. We evaluate our performance with respect to a test set of proteins for which we have known GO mappings. We are currently considering several evaluation measures for extending standard precision and recall metrics from the information retrieval community to the context of ranked annotations in a structured ontology (cf. Kiritchenko et al 2005, Pal and Eisenberg 2005), where the assessment of the correctness of a prediction may depend on the structural relations between the answer and the prediction.

5. REFERENCES

1. The Gene Ontology Consortium. 2000. Gene Ontology: Tool For the Unification of Biology, *Nature Genetics*, 25:1:25-29.

2. Verspoor, K., Cohn, J., Joslyn, C., Mniszewski, S., Rechtsteiner, A., Rocha, L.M., Simas, T. 2005. Protein Annotation as Term Categorization into the Gene Ontology using Word Proximity Networks. *BMC Bioinformatics* 2005 6(suppl 1).

3. Verspoor, K., Cohn, J., Mniszewski, S., Joslyn, C. 2005. Nearest Neighbor Categorization for CASP Function Prediction. In ISMB 2005 poster session. Detroit, MI.

4. http://www.c3.lanl.gov/~joslyn/posoc.html

5. Joslyn, C., Mniszewski, S., Fulmer, A., Heaton, G. (2004). The Gene Ontology Categorizer. *Bioinformatics*, vol. 20, supplement 1, i169-i177.

6. Altschul, S.F., Madden, T.L., et al (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25, 3389-3402.

7. Kiritchenko, S., Matwin, S., Famili, A.F. (2005) "Functional Annotation of Genes Using Hierarchical Text Categorization", to appear in Proc. BioLINK SIG on Text Data Mining.

8. Pal, D., Eisenberg, D. (2005) "Inference of Protein Function from Protein Structure", *Structure* 13.

Automated annotation using functional signatures from structural alignments

Kai Wang, Ram Samudrala*
Computational Genomics Group, Department of Microbiology
University of Washington, Seattle, WA 98195, USA
*To whom correspondence should be addressed: ram@compbio.washington.edu

1. INTRODUCTION

The success of structural genomics initiative requires computational tools that can reliably classify a given protein structure with a known structural fold into the correct functional category. To this end, we have developed a novel method for identifying functional signatures from structural alignments (FSSA) (1). Given an ensemble of proteins sharing the same structural fold, we first perform an all-against-all pairwise structural alignment using the MAMMOTH program (2). For each amino acid residue in a given structure, we calculate the log odds of finding the same local structure in homologues versus structural analogues. For a given protein, the collection of these log odds scores for all residues comprises its functional signature. The functional signatures can be used to interpret the functional importance of each residue or to classify query protein structures into functional categories. We implemented the FSSA method in a web server for automated function prediction, available at <http://protinfo.compbio.washington.edu/fssa>. The Structural Classification of Proteins (SCOP) scheme is a widely used classification method that classifies protein structures into hierarchical levels of class, fold, superfamily and family to embody structural and evolutionary relationships (3). Here we use the SCOP superfamily as a proxy for functional category, and our goal is to confidently classify protein structures with known folds into SCOP superfamilies. Using this scheme, we compare the performance of FSSA with several other sequence and structure homology based function prediction methods (Smith-Waterman, PSI-BLAST, HMM, MAMMOTH and CE).

We strive to solve real-word problems, so we try to make our computational experiments approximate the real-world scenario. There are several marked differences in our evaluation procedures, compared to those used in many publications on function prediction methods. First, although the majority of published methods aim at discriminating homologues from structural analogues (binary decision problem), we aim at assigning a given query sequence into a particular functional category (multi-category classification problem), since it reflects the practical problem we are facing with. Second of all, unlike others that discard functional categories that contain very few sequences, we combine these small categories into a single "OTHER" category. This makes the correct classifications harder, but such procedure does approximate the real situation in automated function prediction. We believe that the results derived from our evaluation procedures can better approximate the situation for functional annotation of structural genomics targets.

2. RESULTS

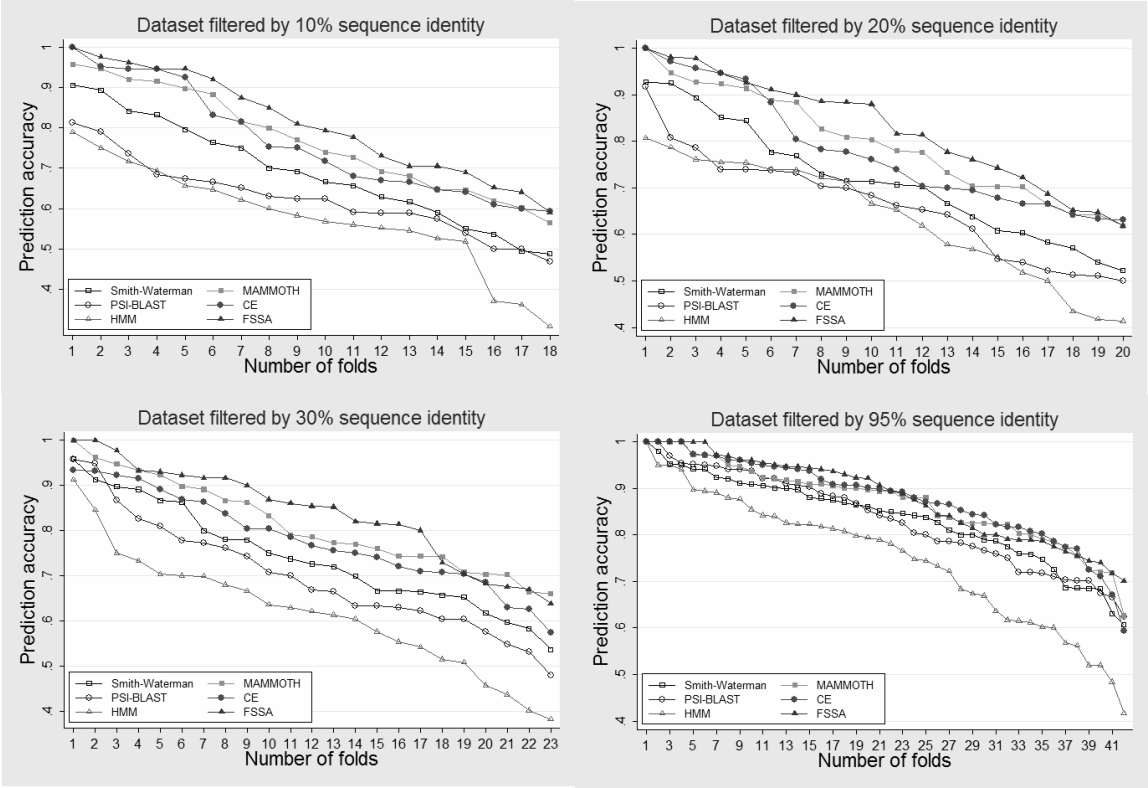
We downloaded the most recent ASTRAL compendium (version 1.67) (4) of the SCOP database for our computational experiments. We discarded all structures containing large missing segments (consecutive C_α atoms >10Å), since structure alignment programs cannot reliably align them. For each SCOP fold we combined all superfamilies with less than eight sequences into a single "OTHER" category. We performed four-fold cross-validation experiments on all SCOP folds that contained at least two superfamilies. In each of the cross-validation experiments, 75% of the domain structures were used as databases and the remaining 25% structures are used as queries. Each query was assigned to the same functional category (SCOP superfamily) as the database sequence with the best homology score (E-value for sequence comparison methods, Z-score for structure comparison methods and log odds score for the FSSA method).

To investigate the correlation between performance and similarity among testing and training sequences, we used four different datasets retrieved from the ASTRAL compendium, representing proteins whose pairwise sequence identity are less than 10%, 20%, 30% and 95%, respectively. For all sequence identity levels, these structural folds in our datasets contain all-alpha, all-beta, alpha/beta, alpha+beta and small proteins, and are good representatives of the fold space. Overall, the FSSA method has the best function classification performance when pairwise sequence identity in the datasets is $\leq$30%, though the differences are subtle between all methods utilizing structural information (Figure 1). Sequence homology based function classification methods perform relatively poorly at low sequence identity levels. Our evaluation

results demonstrated that the FSSA method would be valuable in automated function annotation applications for structural genomics projects, together with other sequence and structure comparison methods.

3. FIGURES

Figure 1. Relative performance of six function classification methods on datasets from the SCOP database that has been filtered by 10%, 20%, 30% and 95% pairwise sequence identity, respectively. For each function classification method, the number of SCOP folds is plotted against the minimum prediction accuracy achieved by that method. The FSSA method has the overall best performance in function classification when sequence identity is $\leq 30\%$.



4. REFERENCES

[1] Wang, K., Samudrala, R. 2005. FSSA: a novel method for identifying functional signatures from structural alignments. *Bioinformatics* in press.

[2] Ortiz, A. R., Strauss, C. E., and Olmea, O. 2002. MAMMOTH (matching molecular models obtained from theory): An automated method for model comparison. *Protein Science* 11:2606-2621.

[3] Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C., Murzin, A. G. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acid Research* 32:D226-D229

[4] Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., Brenner, S. E. 2004 The ASTRAL compendium in 2004. *Nucleic Acids Research* 32:D189-D192.

Prediction and Contradiction of Protein Function using Annotation Rules

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

1. INTRODUCTION

In recent years, high-throughput genome sequencing projects have provided the scientific world with a wealth of decoded protein sequences. Cataloguing the function of these proteins requires a literature curation process in which all available publications about a protein are manually checked. As a consequence of this time-consuming process the proportion of well-characterized proteins compared to the amount of decoded sequences is shrinking.

One way to keep up with the challenge of characterizing the huge amount of protein sequencing data is to develop and to apply automated methods. Two services are provided by the European Bioinformatics Institute (EBI) in order to tackle this challenging task. The RuleBase (2) and Spearmint (3) systems are applied at the EBI on a regular basis in order to predict function and descriptions for the proteins in the TrEMBL section in the UniProt Knowledgebase (1), which would otherwise contain little or no information. A third system, Xanthippe (4) is applied to avoid false positive predictions of the systems and to detect erroneous imports from other databases. Sequences from external users can be submitted to the provided web services that will return an annotated protein predicted by either RuleBase or Spearmint (http://www.ebi.uniprot.org/uniprot-srv/aaTools.do). A description of the systems is given below.

2. METHODS

The predictions give general descriptions like the protein name and more detailed information about a protein's function. All systems predict Swiss-Prot keywords (controlled vocabulary), Swiss-Prot description lines and free-text comments. GO-terms are predicted by one of the systems (Spearmint).

The RuleBase system is based on a manual analysis of well annotated proteins in the Swiss-Prot database. A human expert scans a group of similar protein entries in order to detect correlations between computed sequence patterns like PFAM, PROSITE or SMART and protein annotations like Swiss-Prot keywords. Once such a correlation is detected, e.g. a curator finds out that the keyword *'Serine Protease'* is always annotated for proteins belonging to InterPro IPR000001 when PROSITE pattern PS00134 is present, an annotation rule is created. This rule can be applied to uncharacterized proteins, i.e. if a protein fulfills all conditions of a rule, the annotation item attached to it, is assigned to the protein entry. Although RuleBase proofed to be a high quality annotation prediction system it has some drawbacks in terms of efficiency. A thorough manual investigation of the proteins is necessary, before more rules can be added to the current rule set of 850 rules. Furthermore, maintenance efforts are necessary in order to keep the existing rules up-to-date.

The second prediction system, Spearmint uses a similar approach to the RuleBase system. It overcomes the drawbacks of the latter by using a fully automatic approach. Instead of investigating the proteins manually, a machine learning approach is used. The employed C4.5 decision tree algorithm automatically classifies the proteins of a given training set into positive and negative examples depending on the presence of a particular annotation. The tree generated by the algorithm (Fig. 1) consists of a root node, several condition nodes and leaf nodes. The root and condition nodes represent sequence patterns that can be computed for all sequences by secondary databases like Pfam or ProDom, whereas the leaf nodes hold the annotation. The Spearmint systems extracts rules from those parts of the tree that describe the positive instances of the training-set. About 10 000 rules can be generated this way on a fortnightly basis and can be used to classify uncharacterized sequences with a quality equivalent to RuleBase but with a higher coverage.

Even though the RuleBase and Spearmint systems produce high qualitative annotation (Fig. 2), they introduce erroneous annotation. The systems concentrate on only a small amount of Swiss-Prot data during the learning process, since looking at the whole database would be far too inefficient. Consequently the systems learn incomplete rules that lead to false positive predictions. Xanthippe, a contradiction system, was developed to avoid most of these questionable predictions. It post-processes the annotations of other

systems and marks them, if necessary, as erroneous. For example, if a prediction system predicts the keyword *'Nuclear Protein'* for a bacterial protein, Xanthippe gives a warning that this annotation is wrong. About 700 organism-keyword contradiction rules, like the one mentioned above, were generated manually. In addition about 7000 contradiction rules are generated fortnightly using the decision tree algorithm, described in detail in "*Filtering erroneous protein annotation*" (4).

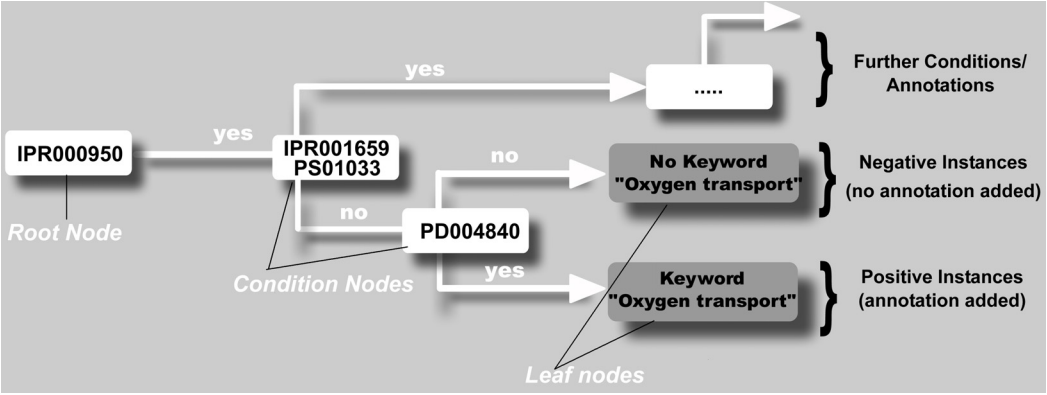


Fig. 1: Decision tree that classifies proteins belonging to IPR000950. The Spearmint prediction rule is: "*Add the Keyword "Oxygen transport", if the protein belongs to IPR000950 (InterPro), but neither to IPR001659 nor to PS01033 (PROSITE) and if it has a hit to PD004840 (ProDom)".*

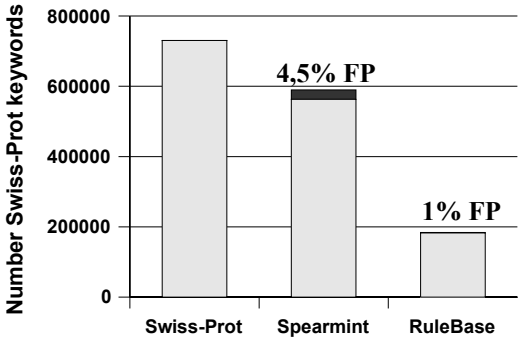


Fig. 2: Quality measurements for Spearmint and RuleBase. Diagram shows true positive (light grey) and false positive keyword predictions (dark grey) of the systems once applied to Swiss-Prot. The contradiction system Xanthippe detects 30% of the Spearmint false positive predictions and 10% of the RuleBase false positive predictions (not shown in the diagram).

3. REFERENCES

1. Apweiler R., Bairoch A., 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., Yeh L.L. 2004. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, 32:D115-D119.

2. Biswas M., O'Rourke J.F., Camon E., Fraser G., Kanapin A., Karavidopoulou Y., Kersey P., Kriventseva E., Mittard V., Mulder N., Phan I., Servant F., Apweiler R. 2002. Applications of InterPro in protein annotation and genome analysis. *Briefings in Bioinformatics 3(3):285-295 (2002)*.

3. Kretschmann E., Fleischmann W., Apweiler R. 2001. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, 17:920-926.

4. Wieser D., Kretschmann E., Apweiler R. 2004. Filtering erroneous protein annotation. *Bioinformatics*, 20**:**i342-i347.

Posters

Poster abstracts can be found at http://ffas.burnham.org/AFP

Using Physicochemical Properties to Identify Protein Functional Sites

Mike P. Liang, Shirley Wu, Russ B. Altman*
Department of Genetics, Stanford University,
300 Pasteur Drive L301 MC:5120, Stanford, CA 94305-5120, USA
*To whom correspondence should be addressed: russ.altman@stanford.edu

Function Prediction and Integrated Mining of Sequence and Structure Features with Unison

Reece Hart[1,2]*, Kiran Mukhyala[1]
Genentech, Inc., Departments of Bioinformatics[1] and Protein Engineering[2],
1 DNA Way, MS-93, South San Francisco, CA 94080, USA
*To whom correspondence should be addressed: rkh@gene.com

Learning Classifiers for Assigning Protein Sequences to Gene Ontology Functional Families

Carson Andorf[1,2,5], Adrian Silvescu[1,2,5], Drena Dobbs[3,4,5], Vasant Honavar[1,2,4,5*]
[1]Artificial Intelligence Laboratory, [2]Department of Computer Science
[3]Department of Genetics, Development, and Cellular Biology
[4]Bioinformatics and Computational Biology Graduate Program
[5]Computational Intelligence, Learning, and Discovery Program
Iowa State University, Ames, IA 50010, USA[1]
*To whom correspondence should be addressed: honavar@cs.iastate.edu

Hidden Markov Models Hierarchical Classification for *ab-initio* Prediction of Protein Subcellular Localization

Richard Hugues[*,1,] Mucchielli Marie-Hélène[2], Prum Bernard[1], Képès François[1]
[1]Laboratoire Statistique et Génomes, place des Terrasses, 91000 Evry, France.
[2]Centre de Génétique Moléculaire, Gif-sur-Yvette.
*To whom correspondence should be addressed: richard@genopole.cnrs.fr

A generalized strategy for reconstructing genome-wide functional interaction maps using a naïve Bayesian framework: An application to the *P. falciparum* genome

Shailesh V. Date & Christian Stoeckert, Jr.
Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104
svdate@pcbi.upenn.edu

What Can We Learn About Genome From 2-D DNA Walk?

S.A.Larionov*, A.Loskutov, E.V.Ryadchenko
Physics Faculty, Moscow State University, Moscow 119992, Russia
*To whom correspondence should be addressed: Serglarionov@yandex.ru

Exploitation of Protein Structural Information for Divergent Protein Function Prediction with Machine Learning Approach[a]

Ying Lin[1, *], John Case[1], Lappoon Rupert Tang[2], Hsing-Kuo Kenneth Pao[3], Joan Burnside[4]
[1]Department of Computer and Informatics Sciences, University of Delaware, Newark, DE 19716, U.S.A.
[2]Department of Computer Sciences, University of Texas at Brownsville, Brownsville, TX, 78520,U.S.A.
[3]Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, 106, Taiwan.
[4]Department of Animal and Food Sciences, University of Delaware, Newark, DE 19717, U.S.A.
*To whom correspondence should be addressed: ylin@cis.udel.edu

Automated Functional Inference of Enzyme Mutants Utilizing a Four-Body Statistical Potential

Majid Masso*, Iosif I. Vaisman
Bioinformatics and Computational Biology, School of Computational Sciences,
George Mason University, 10900 University Blvd., MSN 5B3, Manassas, VA 20110.
*To whom correspondence should be addressed: mmasso@gmu.edu

Conservation of Residues in Structure-Based Functional Site Predictions

Huyuan Yang, Mary Jo Ondrechen*
Department of Chemistry and Chemical Biology and Institute for Complex Scientific Software,
Northeastern University, Boston, MA 02115, USA
*To whom correspondence should be addressed: mjo@neu.edu

Site Prediction for Protein Structures that Undergo Conformational Change upon Ligand Binding

Yongli Gao, Leonel F. Murga, Mary Jo Ondrechen *
Department of Chemistry & Chemical Biology and Institute for Complex Scientific Software, Northeastern University, Boston, MA 02115 USA
*To whom correspondence should be addressed: mjo@neu.edu

The Proteome Analyst Suite of Automated Function Prediction Tools

Poulin B.*, Szafron D., Lu P., Greiner R., Wishart D.S.,
Eisner R., Fyshe A., Pearcy B., and Pireddu L.
Department of Computing Science, University of Alberta, 221 Athabasca Hall, T6G 2E8, Canada
*To whom correspondence should be addressed: poulin@cs.ualberta.ca

Experimental Evidence for the Functional Importance of Residues Predicted by THEMATICS

Heather Brodkin[a], Amy C. Milne[b], Mary Jo Ondrechen[a] and Dagmar Ringe[b]*
[a]Department of Chemistry and Chemical Biology and Institute for Complex Scientific Software, Northeastern University, Boston, MA, 02115, USA, and [b]Department of Biochemistry, Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA 02454-9110, USA
*To whom correspondence should be addressed: ringe@brandeis.edu

FunCat functional assignment by Belief Propagation inference and feature integration

Dmitrij Surmeli[a]*, Oliver Ratmann[b], Igor Tetko[a] and Hans-Werner Mewes[a]
[a]Institute for Bioinformatics, GSF GmbH, 85764 Nueherberg, Germany
[b]Universite di Lecce, 73100 Lecce, Italy
*To whom correspondence should be addressed: dmitrij.surmerli@gsf.de

MIPS-BFAB – an Example of Automatic Sequence Annotation Benchmarking Resource

Igor V. Tetko,*,1 Andreas Ruepp, Hans-Werner Mewes1,2
1-GSF National Research Center for Environment and Health, Institute for Bioinformatics, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany, 2- Department of Genome-Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, 85350 Freising, Germany,
*i.tetko@gsf.de

Identification of Protein Active Sites Using Theoretical Microscopic Titration Curves and Support Vector Machines

Wenxu Tong[b,c], Mary Jo Ondrechen[a,c], Ronald J. Williams*[b,c]
Department of Chemistry and Chemical Biology[a],
College of Computer and Information Science[b],
and Institute for Complex Scientific Software[c],
Northeastern University
Boston, MA 02115 USA
*To whom correspondence should be addressed: rjw@ccs.neu.edu

Combining ZDOCKpro with Evolutionary Trace to improve the protein-protein docking results

Lisa Yan
Accelrys, 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA
*To whom correspondence should be addressed: lly@accelrys.com