Automated Function Prediction SIG 2012

Organizers: Iddo Friedberg and Predrag Radivojac

08:30	08:45	Welcome	Iddo Friedberg and Predrag Radivojac	
		Session 1: Evolution used in Function Prediction		
08:45	09:30	Keynote: Phylogenomic Approaches to Improving Functional Prediction	Jonathan Eisen	UC Davis
09:30	09:50	FAT-CAT (Fast Approximate Tree ClassificATion): A Scalable Method for (Meta)Genome Functional and Taxonomic Classification	Kimmen Sjolander	UC Berkeley
09:50	10:15	Annotation of Short Proteins in the Expanding Universe of Protein Sequences from Insects and Crustaceans	Michal Linial	Hebrew University of Jerusalem
10:15	10:45	Coffee		
		Session 2: Structures!		
10:45	11:10	Protein-ligand binding prediction with I-TASSER structure assembly, BioLiP database construction and COFACTOR template identification	Yang Zhang	University of Michigan
11:10	11:30	STOP using GO	Sean Mooney	Buck Institute
11:30	11:55	Evolution of function in the alkaline phosphatase superfamily	Alan E. Barber Ii	UC San Francisco
11:55	12:20	Selection of Targets for Function and Structure Determination in the Isoprenoid Synthase Superfamily	Daniel Almonacid	UC San Francisco
12:20	13:30	Lunch		
13:30	14:30	posters		
		Session 3: Getting creative.		
14:30	14:55	Can we go beyond sequence similarity to predict protein function ?	Stephano Toppo	Università degli studi di Padova
14:55	15:15	Computational Function Predictions for Moonlighting Proteins	Daisuke Kihara	Purdue University
15:15	15:45	Coffee		
		Session 4: Critique of Protein Function Prediction		
15:45	16:10	Lessons from the first critical assessment of functional annotation	Jesse Gillis	University of British Columbia
16:10	16:35	Sources of Experimental Function Annotation in UniProt-GOA, Implications for Function Prediction	Alexandra Schnoes	UC San Francisco
16:35	17:00	Information theory based metrics for the evaluation of GO term annotations	Wyatt Clark	Indiana University, Bloomington
17:00	17:15	Break		
17:15	17:40	Predicting Tissue Specificity from Protein Sequence	Sivan Goren	Bar Ilan University
17:40	18:00	“dcGO”: a Domain-Centric Gene Ontology Predictor for Functional Genomics	Julian Gough	Bristol University, UK
18:00	18:25	Flexible Graphlet Kernels for Functional Residue Prediction in Protein Structures	Jose Lugo-Martinez	Indiana University, Bloomington

FAT-CAT (Fast Approximate Tree ClassificATion): A Scalable Method for (Meta)Genome Functional and Taxonomic Classification

Bushra Samad, Jonathan Dobbie, Kimmen Sjölander*
University of California Berkeley, Berkeley CA 94720 USA
*kimmen@berkeley.edu

1. INTRODUCTION

We present the FAT-CAT (Fast Approximate Tree ClassificATion) algorithm to automate the functional annotation of genomes and the simultaneous functional and taxonomic annotation of metagenome sequences. FAT-CAT makes use of pre-calculated phylogenetic trees for multi-domain architectures and for Pfam domains in the PhyloFacts 3.0 database (1). PhyloFacts is a phylogenomic encyclopedia of gene families across the Tree of Life; it contains >7.3M protein sequences from >99K taxa (including strains) drawn from the UniProt database. Our PhyloFacts library construction pipeline uses a suite of software tools to cluster and align proteins, build phylogenetic trees, retrieve various types of data and construct hidden Markov models. Homologs for Pfam domain families are retrieved using Jackhmmer (2) and we use the FlowerPower algorithm to cluster proteins into multi-domain architectures (3). We construct multiple sequence alignments using MAFFT (4) and estimate phylogenetic trees using FastTree (5). We retrieve experimental and other annotation data from a variety of resources, including the Enzyme Commission (EC), Gene Ontology (GO), BioCyc and SwissProt and then overlay these data onto phylogenetic trees for visual inspection by users. Most PhyloFacts families include duplication events (i.e., paralogous groups) and trees for Pfam domains typically also include domain architecture rearrangements; in consequence, most PhyloFacts families span many divergent functions and structures. Orthologs are identified using the PHOG algorithm (6) and functional subfamilies are identified using SCI-PHY (7).

The PhyloFacts library construction pipeline is computationally expensive; both CPU and disk space constraints prevent the use of this pipeline every time a new genome is sequenced. The FAT-CAT system is designed to provide rapid and highly specific functional sub-classification of novel sequences without the computational burden of library construction. FAT-CAT uses HMMs at internal nodes of PhyloFacts trees, which are annotated with the observed functions (e.g., EC numbers, GO annotations, etc.) of sequences within the trees. Classification of sequences to these HMMs allows us to predict function (and possibly taxonomic origin). FAT-CAT is related conceptually to both subfamily HMMs (7) and to the TreeHMM method of Goldstein and colleagues (8). However, the Goldstein TreeHMM method was designed to improve remote homolog detection whereas subfamily HMMs and FAT-CAT are designed for the purpose of functional sub-classification.

Relative to subfamily HMMs, FAT-CAT provides increased flexibility in the classification of sequences. Logistic regression of subfamily HMM scores allows us to differentiate between sequences that can be assigned to SCI-PHY subfamilies from those that represent novel subtypes. Classification of a sequence as a probable novel subtype effectively places the sequence at the root of the tree, where functional divergence may be extreme, and provides minimal information about the protein's possible function. However, the availability of HMMs at all internal nodes in the tree allows FAT-CAT to classify a sequence to a point higher in the tree (i.e., toward the root) providing some clues to its function. In other cases, a placement deeper *within* a subfamily may be possible, and allow both a more specific prediction of function as well as the taxonomic origin. The latter are particularly important in the case of metagenome data analysis.

We are exploring different techniques to improve FAT-CAT scalability to large datasets while maintaining high precision. First, we use the HMMER 3.0 suite (2) which has been optimized for speed. We then reduce the number of HMM scores required using a two-step protocol. First, we select families for sub-classification by scoring query sequences against family HMMs (these correspond to HMMs located at the root nodes of PhyloFacts trees). Even with the almost 100,000 family HMMs in PhyloFacts, this initial step takes under a minute on average. Families with significant scores are then selected for phylogenetic placement using FAT-CAT. HMMER's super-fast hmmscan software makes a brute-force approach feasible (i.e., scoring the query sequence against all HMMs in the tree). We are also developing a tree-traversal approach. Tree traversal recursively traces a path from the root to a leaf, starting at the root node and scoring the query against the HMMs located at child nodes. We then follow the edge to the child node

corresponding to the HMM giving the strongest score. The process is repeated until a leaf node is reached. The HMM giving the query the most significant score on that path is identified, and the corresponding subtree node is used to derive a functional (and perhaps taxonomic) annotation. Using tree traversal reduces the number of HMM scores required to place a sequence in a balanced binary tree of K sequences to only O(log K) making phylogenetic placement efficient.

Our preliminary data shows FAT-CAT is competitive with the top-ranked methods in phylogenetic placement (e.g. EPA (9)). We will also report results in functional classification of sequences using leave-1-out experiments on the manually curated Structure Function Linkage Database (SFLD) benchmark (10) and on simulated metagenome datasets.

2. REFERENCES

1. Krishnamurthy N, Brown DP, Kirshner D, Sjölander K. 2006. PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. *Genome biology* 7(9):R83
2. Eddy S R 2009. A New Generation of Homology Search Tools Based On Probabilistic Inference. *Genome Informatics,* 23:205-211
3. Krishnamurthy N, Brown DP, Sjölander K. 2007. FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. *BMC Evolutionary Biology* 7 Suppl 1:S12
4. Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics* 9(4):286-298.
5. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* 5(3): e9490
6. Datta R.S, Meacham C, Samad B, Neyer C, Sjölander K. 2009. Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Research* 37(Web Server issue):W84-89
7. Brown DP, Krishnamurthy N, Sjölander K. 2007. Automated protein subfamily identification and classification. *PLoS Computational Biology* 3(8):e160
8. Qian B, Goldstein RA. 2003. Detecting distant homologs using phylogenetic tree-based HMMs. *Proteins* 52(3):446-453
9. Berger SA, Krompaß D, Stamatakis A. 2011. Performance, Accuracy and Web-Server for Evolutionary Placement of Short Sequence Reads under maximum-likelihood. *Systematic Biology* 60(3):291-302
10. Pegg SC, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC. 2006. Leveraging Enzyme Structure-Function Relationships for Functional Inference and Experimental Design: The Structure-Function Linkage Database. *Biochemistry* 45:2545-2555

Annotation of Short Proteins in the Expanding Universe of Protein Sequences from Insects and Crustaceans

Nadav Rappoport[1] and Michal Linial[2]
[1]School of Computer Science and Engineering, [2]Deptartment of Biological Chemistry, Institute of Life Sciences, The Sudarsky Center for Computational Biology, The Hebrew University of Jerusalem, Israel

*To whom correspondence should be addressed: nadavrap@cs.huji.ac.il

1. INTRODUCTION

The overwhelming increase in protein sequences made available by current technology has impacted our ability to annotate the emerging collection of sequenced genomes. UniProt is the main archive for proteins with over 21 million sequences (2012, www.uniprot.org). Multicellular organisms, including metazoa and plants, account for almost 4 million sequences. In reality, 40% of this set (1.6 millions) are fragmented ORFs with unspecified start or stop codons. Identifying full length ORFs and their functions remain major obstacles. Functional annotation task lags behind for certain proteins types, such as short proteins.

Most genome annotation tools fail to identify short proteins. Many of these sequences are associated with low statistical scores coming from search engines (Profile-Profile, Psi-Blast). Specifically, the score for a secure identification of a gene from deep sequencing or mass spectrometry (MS) is often below a predetermined threshold. From a structural perspective, the short proteins do not have a classical hydrophobic core. The rigid scaffold of their backbone is mostly a result of a coordinated formation of disulfide bonds. Importantly, short proteins are rapidly evolving. Evolutionary processes such as duplication, positive selection and convergence evolution drive the evolution of short proteins. Therefore, a major task in the expanding universe of protein sequences is the assignment of functions to short proteins.

Evidence from comparative genomics from model organisms (e.g., a dozen species from Drosophilae, ModEncode project) led to the discovery of previously overlooked putative proteins [1]. However, the majority of short ORFs remain poorly annotated. The coverage of Arthropods and specifically the insects had been expanded with genomes from flies (12), mosquitoes (3), bees (2), ants (6), butterflies (2) and others including beetle, wasp and tick and lice.

2. METHODS

We undertake the task of functional annotations assignment for short proteins from Arthropods. Incorporation and merging three major tools and concepts improved the discovery rate and the success in functional annotation:

(i) *Stable clusters from ProtoNet tree.* A stringent charting of the protein sequence space by ProtoNet algorithm serves as a scaffold for grouping short proteins.

(ii) *Signal Peptide (SP) containing proteins*. We focus of protein sequences that undergo post-translational modifications (PTMs) following translocation to the endoplasmic reticulum (ER). Most of the short secreted proteins have characteristic SP.

(iii) *ClanTox predictor for short proteins.* ClanTox is a machine-learning tool that identifies properties such as the structural compactness and the cysteine-rich backbones in animal proteins. As such, it is a powerful tool for identifying compact, disulfide bonded proteins.

For rigorously testing the annotation potential of short proteins, we constructed a new platform that is based on the ProtoNet tree concept [2]. To avoid the overwhelming statistical noise from millions of proteins (50% are from bacteria), we focused on available complete sequenced genomes from insects. Our database comprises of 300,000 protein sequences that were collected from genomic centres. The platform, called ProtoBug, is a genome-based hierarchical family tree. To avoid a bias due to partial sequences from Drosophilae [1], we only included fully sequenced species. We thus expanded the phylogenetic breath among insects. In addition, we included the Crustacean *Daphnia pulex* proteome [3] as outgroup. The Daphnia's proteome covers 30,000 sequences of which 98% are named "uncharacterized".

3. RESULTS

The database of about 18 'complete' proteomes provides an opportunity to closely examine related proteomes, while avoiding the statistical noise (e.g., Blast E-value). Only 3.5% of sequences in ProtoBug have experimental support (transcripts and proteomics), but evidence is available to <2% among short sequences (≤100 aa).

ProtoBug takes advantage of the high quality platform of ProtoNet [2]. The main power of ProtoBug tree constructing is the use of an unsupervised, bottom-up agglomerative averaging protocol that outperforms naïve algorithms. Our results focus on the analysis of exhaustively studied and poorly studied proteomes (fruitfly and Daphnia, respectively). While 86% of the proteome of Daphnia was successfully annotated using ProtoNet approach (mainly via GO and InterPro) [4], we were able to assign an informative function to only 73% of the short proteome when we used external annotation resources.

A genomic perspective indicates that each genome differs by variety of parameters. For example, 60% of the genomes of the fruitfly and the Daphnia contain paralogs. However, only in Daphnia we detect a large fraction of paralogs with extremely low divergence (Fig. 1A). This finding is in accord with the notion that the evolution rate varied drastically among the Arthropod genomes.

ProtoBug resource contains 76,578 short sequences (≤150 aa) and half of them are shorter than 100 aa. About 50% of ProtoBug's sequences lack external annotation (as they originated from genomic centres). In this research, we activated predictors for the presence of transmembrane domains (TMDs) and Signal Peptide (SP). We were able to assign function to short proteins from the secreted proteomes of Arthropods. The main functional groups include extracellular proteases, lipases, toxins (such as the spider venom) and neuropeptides. However, the function of many of these sequences is still unknown.

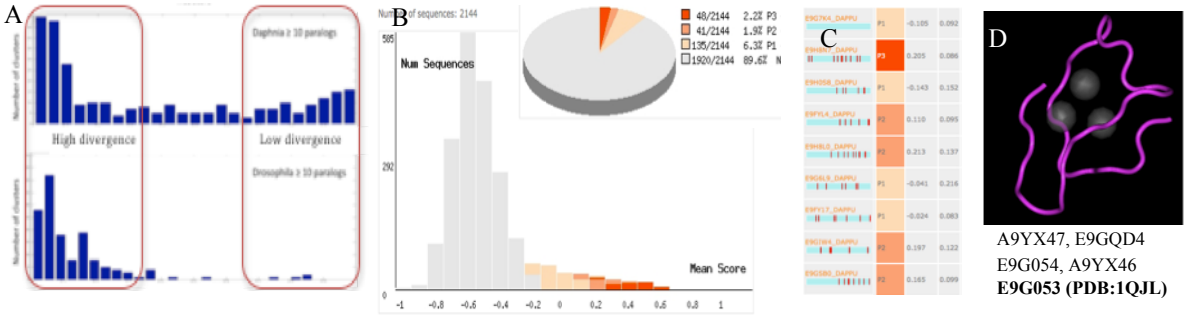


Fig. 1. *Annotation inference protocol and examples.* ***(A)*** *Divergence among clusters with Drosophila and Daphnia paralogs.* ***(B)*** *Results from the short proteome of Drosophila by Clantox.* ***(C)*** *A partial list of Daphnia's predictions as Toxin-like. The distribution of the Cys along the sequence is shown as well as the prediction significant (P3 to P1).* ***(D)*** *Homology modelling of 5 Daphnia's proteins (E9G053, uncharacterized). All 5 proteins are modelled as metallothionein. The NMR structure (PDB:1QJL) from Sea Urchin was used to model the Daphnia's proteins.*

Among the 3550 and 2145 short proteins from Daphnia and the *D. melanogaster* (≤100 aa, excluded fragments), we identified at high probability, 24 proteins in Daphnia and 89 in Drosophila as toxin-like proteins (Fig. 1B, marked P3 and P2). By lowering the threshold, we extended the discovery to 224 (Drosophila) and 83 (Daphnia) candidates. The definition of toxin-like proteins according to ClanTox platform [5] covers a wide range of functions including proteases, lipases, ion channel modulators (as in spider venom) and numerous peptides acting for defense against fungi and other pathogens. Among the fruitfly predictions, we have successfully annotated a family of over 30 'Defensin-like' proteins with γ-purothionin fold. Interestingly, no homologues were detected in any other insects. The structure of these proteins (representatives are Dro1-6, Dromycin) is similar to scorpion toxins. In the case of the Daphnia short proteome, we identified paralogous proteins that act as metallothionein (Fig. 1D). We postulate that these proteins are activated in the process of heavy metal toxicity. We conclude that while the annotation of large number of proteins from complete proteomes is a challenging task, a combination of functional predictors (ClanTox, SP) and clustering (ProtoBug) is powerful towards this task. Many of the short proteins act in pathogen defense mechanism and act for coping with extreme environmental conditions.

4. REFERECES

1. *M. Muers, Nat Rev Genet* ***12*** *(2), 80 (2011).*
2. *N. Rappoport, S. Karsenty, A. Stern et al., Nucleic Acids Res* ***40*** *(Database issue), D313 (2011).*
3. *J. K. Colbourne, M. E. Pfrender, D. Gilbert et al., Science* ***331*** *(6017), 555 (2011).*
4. *N. Rappoport and M. Linial, BMC Bioinformatics (CAFA/AFP Special issue) (2012).*
5. *G. Naamati, M. Askenazi, and M. Linial, Nucleic Acids Res* ***37*** *(Web Server issue), W363 (2009).*

Protein-ligand binding prediction with I-TASSER structure assembly, BioLiP database construction and COFACTOR template identification

Jianyi Yang, Ambrish Roy and Yang Zhang*
Department of Computational Medicine and Bioinformatics,
University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA

*To whom correspondence should be addressed: zhng@umich.edu

1. INTRODUCTION

Most proteins perform their biological functions through interacting with other molecules. Therefore, prediction of the ligand-protein interactions is an important problem in computational protein function annotations. In the recent community-wide CASP9 experiment (1), it was shown that the most accurate approach for ligand binding site prediction (LBSP) is template-based modeling, which makes use of the ligand binding information from homologous structures in the Protein Data Bank (PDB) (2). Essentially, template-based LBSP is a two-step procedure. The first step is to construct correct models of the query sequence, which are structurally close to the native structure of the query protein. In the second step, homologous templates in database of known ligand-protein interactions are identified and used for predicting ligand binding site in the query protein. In our approach, we have used the well-established I-TASSER algorithm for protein structure prediction (3-4). Here, we outline our recent efforts to address the issues in the second step, which includes the construction of the BioLiP database for biologically relevant ligand-protein interactions and the development of COFACTOR algorithm for structure-based template identification and ligand binding site prediction.

2. METHODS

BioLiP database. Examining the biological relevance of the ligands present in PDB files is a non-trivial problem, as a large fraction of the co-crystallized small molecules are usually additives used for solving protein structures (5-6). BioLiP (7) is a new database for **Bio**logically relevant **Li**gand-**P**rotein interactions. To decide the biological relevance of a ligand, we developed an automated composite approach for judging the biological relevance of each ligand present in a PDB file. Due to page limit, the details are presented at http://zhanglab.ccmb.med.umich.edu/BioLiP/about.html. This approach is used to build a comprehensive database BioLiP, which will be used later for LBSP.

COFACTOR algorithm. COFACTOR (8-9) is a new algorithm for structure-based function annotations, which is illustrated in Figure 1. It first identifies template proteins of similar fold/topology by scanning the query structure models through the BioLiP database. The alignments are generated by TM-align (10) and the structural similarity is evaluated using TM-score (11). In the second step, the ligand binding sites of template proteins with the highest TM-score are matched with local structural motifs of query protein, where the query motifs are generated by gleaning conserved residues from the query structure. Starting from a spatial sphere centered at the template binding site, the binding locations on the query protein are identified and further refined using iterative dynamic programming superposition. Finally, consensus LBSP is deduced based on local structural matches to multiple template proteins.

3. RESULTS

A comprehensive database BioLiP has been constructed with the composite automated approach, as described in Section 2. As of March 28, 2012, BioLiP contains 202,564 ligand-protein interaction sites. Four distinct searching engines are provided to search through BioLiP database. The database is freely accessible and available for download at http://zhanglab.ccmb.med.umich.edu/BioLiP/.

The COFACTOR algorithm has been extensively benchmarked on a non-redundant set of 501 proteins (9) (http://zhanglab.ccmb.med.umich.edu/COFACTOR/benchmark/). After excluding all homologous templates with sequence identity to the query of >30%, the average Matthews correlation coefficient,

Precision, and Recall of ligand binding site prediction are 0.58, 0.72, and 0.51, respectively, which are significantly higher than other state-of-the-art methods. The algorithm was recently tested in the community-wide CASP9 blind experiment. As shown in Figure 1, COFACTOR (i.e., 'I-TASSER FN' in CASP9) was ranked as the No 1 method for automated function prediction with a significant margin to all other methods in the field (1). The COFACTOR server is freely available at http://zhanglab.ccmb.med.umich.edu/COFACTOR/.

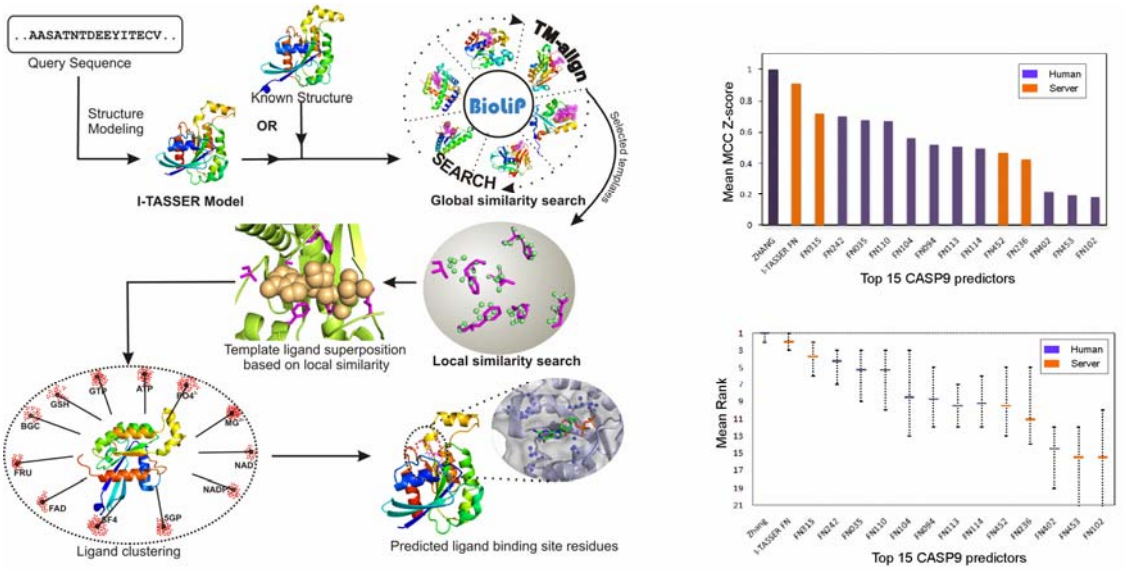


Figure 1 Flowchart of COFACTOR algorithm (left panel) and performance of the top 15 groups in binding site prediction category (right panel, data taken from the CASP9 assessors (1)). The top two groups ('Zhang' as human group and 'I-TASSER FN' as server group) used COFACTOR to predict the binding site residues in protein structures obtained from I-TASSER predictions.

4. REFERENCES

1. Schmidt T, Haas J, Gallo Cassarino T, Schwede T. 2011. Assessment of ligand-binding residue predictions in CASP9. *Proteins* 79: 126-36
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* 28: 235-42
3. Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 5: 725-38
4. Xu D, Zhang J, Roy A, Zhang Y. 2011. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins* 79: 147-60
5. Dessailly BH, Lensink MF, Orengo CA, Wodak SJ. 2008. LigASite--a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res* 36: D667-D73
6. Lopez G, Valencia A, Tress M. 2007. FireDB--a database of functionally important residues from proteins of known structure. *Nucleic Acids Res* 35: D219-D23
7. Yang J, Roy A, Zhang Y. 2012. BioLiP: a comprehensive database for biologically relevant ligand-protein interactions. *In preparation*
8. Roy A, Yang J, Zhang Y. 2012. COFACTOR: An accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res* In press
9. Roy A, Zhang Y. 2012. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure*: In press
10. Zhang Y, Skolnick J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic. Acids Res.* 33: 2302-9
11. Zhang Y, Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins* 57: 702-10

Evolution of function in the alkaline phosphatase superfamily

Alan E. Barber II* (1), Jonathan K. Lassila (2), Helen I. Weirsma-Koch (2), Michael A. Hicks (1), Daniel Herschlag (2), Patricia C. Babbitt (1)
1. Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA
2. Department of Biochemistry, Stanford University, Stanford, CA
*To whom correspondence should be addressed: alan.barber@ucsf.edu

1. INTRODUCTION

Mechanistically diverse enzyme superfamilies are composed of evolutionarily related enzymes that share common mechanistic features yet catalyze different chemical reactions. The study of the evolution of these types of superfamilies provides a unique opportunity to understand how nature has modified specific catalytic scaffolds to catalyze numerous enzymatic reactions. The alkaline phosphatase (AP) superfamily is a good model for the use of evolutionary relationships to understand specialization of enzymatic function. This is because its founding member, AP, represents a prototypic phosphoryl transfer catalyst and is among the most well characterized enzymes. Homologues of AP share a structural core with similar active site and catalyze a range of phosphoryl and sulfuryl transfer reactions including phosphatases, sulfatases, phosphodiesterases and phosphomutases. (Figure 1A, 1) Additionally, catalytic promiscuity is widespread in the AP superfamily, with secondary activities measured for enzymes with primary phosphatase, sulfatase, phosphodiesterase, phosphomutase, and phosphonoacetate hydrolysis activities (2). This promiscuity complicates computational functional prediction and provides an experimental and computational system to understand enzyme evolution.

Protein similarity networks (PSNs) are graphical representations of sequence, structural, and other types of similarities among a group of proteins in which pairwise all-by-all similarity connections are calculated. Nodes are used to represent one or more protein sequences or structures and edges drawn between two nodes represent some measure of their similarity. Mapping biological information to network nodes or edges enables hypothesis creation about sequence-structure-function relationships across entire sets of related proteins. (3)

We present an investigation of the AP superfamily using PSNs to hypothesize an evolutionary model, which we evaluate with phylogenetic analysis. This model demonstrates the elaboration of an ancient minimal precursor protein into the various known modern day functions resulting from complex mechanisms evolving both divergently and convergently (from intermediate ancestsors in the superfamily tree) functions. This study demonstrates many pitfalls for automated function prediction in a mechanistically diverse enzyme superfamily along with some suggestions for addressing these issues.

2. RESULTS AND DISCUSSION

Guided by conserved structural elements of the fold, we constructed sequence similarity networks to establish a global view of similarity relationships across the AP superfamily (data not shown). Using statistically significant scoring thresholds for creating edges between nodes at which biologically relevant groups begin to distinguish themselves (E-value < 8E-14), a set of nucleotide pyrophosphatase/phosphodiesterase-like (NPP-like) enzymes were identified as forming a hub that connects multiple subgroups within the superfamily. This result indicates that these NPP-like enzymes have the most general sequence properties of the larger superfamily. Additionally, based on an analysis of the topological variations within the AP superfamily, we identified these central NPP-like enzymes as being minimal; that is, these proteins lack domain insertions relative to the other subgroups of the AP superfamily. Based on this evidence, we hypothesize that minimal NPP-like enzymes may represent a modern-day version of an ancestral enzyme.

To evaluate the model, we constructed a phylogenetic tree of representative members of the AP superfamily. Because of the extreme divergence of some of the superfamily members, a well-resolved tree for the entire superfamily could not generated; instead, we focused on a more tractable subset representing the majority of members of the AP superfamily, those using a conserved two-metal-ion site (data not shown). Several central nodes in this tree were resolved with > 0.95 confidence, provding some validation of our evolutionary model (Figure 1B).

This model demonstrates the difficultly in using solely sequence similarity to make functional inferences

for such divergent proteins. Monoesterases of similar function (alkaline phosphatases and NPP-like monoesterases) are found on two distinct branches of the tree and use distinct, apparently independently evolved strategies to catalyze the reaction. More closely related to these than they are to each other, other seemingly more divergent enzymes in terms of function (phosphonoacetate hydrolase to NPP-like monoesterases and 2,3-bisphosphoglycerate-independent phosphoglycerate mutase to APs) are of the same clade but have evolved major structural variations by mutation or domain insertion.
We used the diesterases in the AP superfamily as a model to investigate how sequence and structure features can be mapped to specific catalytic functions. We identified active site residues and clustered the sequences based on sequence similarity (data not shown). By mapping co-evolution of these functional residues onto the PSN, we were able to determine functional boundaries that correspond with known experimentally determined functions and identify, for further characterization, sequences with likely novel function.

3. CONCLUSIONS

We will present a model for the evolution for the AP superfamily in which a minimal precursor protein evolved to produce the diverse set of sequences, structures and functions seen in modern day enzymes This was achieved using several types of structural variation, including loop insertions, domain insertions, loop deletions, point mutations and variations in metal binding constellations at the active sites. This analysis demonstrates how the core AP fold has evolved through both convergent and divergent processes to stabilize the transition states of a number of phosphoryl and sulfuryl transfer reactions and suggests strategies for correctly annotating function in this and other mechanistically diverse enzyme superfamilies.

4. FIGURES

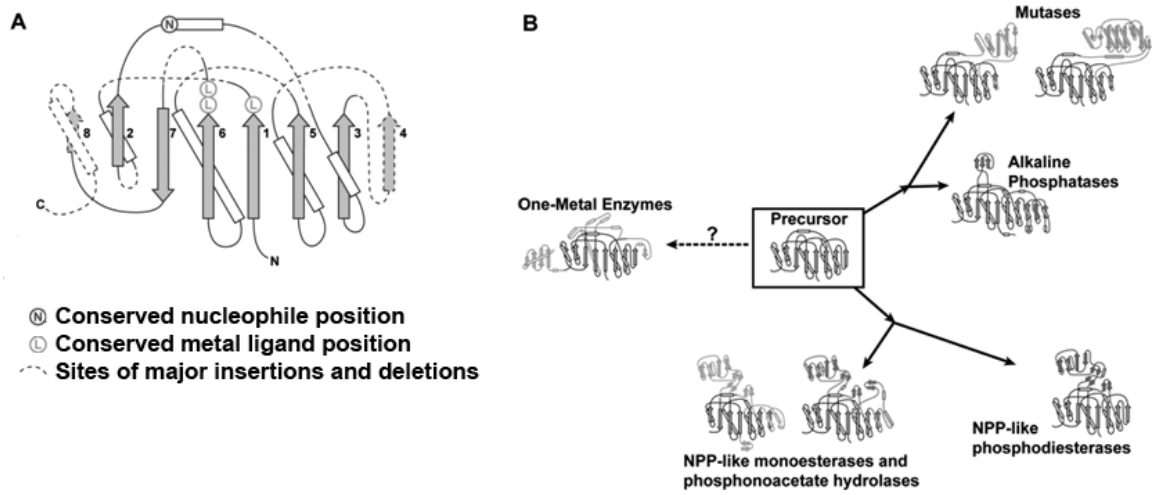


Figure 1. Model of evolution of functional diversity in the AP superfamily. A) Topology diagram of AP superfamily members with universally conserved nucleophile and metal ligand positions indicated by N and L respectively. Major insertions occur in different places in different subgroups. Sites of insertions are indicated with dashed lines. B) Cartoon depiction of our evolutionary model. Fold fopologies of representative experimentally determined functions are shown.

5. REFERENCES

1. Galperin, M. Y., Bairoch, A. and Koonin, E. V. 1998. A superfamily of metalloenzymes unifies phosphopentomutase and cofactor-independent phosphoglycerate mutase with alkaline phosphatases and sulfatases. *Protein Science* 7:1829-1835.

2. Jonas, S. & Hollfelder, F. 2009. Mapping catalytic promiscuity in the alkaline phosphatase superfamily. *Pure and Applied Chemistry* 81:731-742.

3. Brown, S.D. and Babbitt, P.C. 2011. Inference of functional properties from large-scale analysis of enzyme superfamilies. *The Journal of Biological Chemistry* 287:35-42.

Selection of Targets for Function and Structure Determination in the Isoprenoid Synthase Superfamily

Daniel E. Almonacid*, Alexandra M. Schnoes, Patricia C. Babbitt
Department of Bioengineering and Therapeutic Sciences, and California Institute for Quantitative Biosciences, University of California San Francisco, 1700 4th Street, CA 94158, MC 2550, USA
*To whom correspondence should be addressed: daniel.almonacid@ucsf.edu

1. INTRODUCTION

As sequencing expands, the proportion of gene products that can be experimentally characterized becomes vanishingly small. Organization of homologous proteins of known and unknown structures and functions into superfamilies provides a context that can inform function prediction for unknowns (1). The Enzyme Function Initiative (EFI) is a computational/experimental partnership for developing a large-scale sequence/structure-based strategy for functional inference of enzymes from functionally diverse superfamilies (2). Using as an example the isoprenoid synthase (IS) superfamily, we discuss how the EFI bioinformatics core organizes the known structure and function information for superfamily members, and how this knowledge is used to guide selection of targets for experimental and computational characterization that can aid in functional inference for many unknowns.

2. THE ISOPRENOID SYNTHASE SUPERFAMILY

Based on sequence and structural information, we identified 9,925 enzyme sequences having high sequence similarity to a group of α-helical bundle fold enzymes that had been characterized to catalyze C-C bond forming reactions. All these reactions are initiated by Mg^{2+}-assisted dissociation of a pyrophosphate moiety from an allylic diphosphate substrate with concomitant generation of a carbocation intermediate, which can then follow a number of different mechanistic routes. What it is interesting about this superfamily of enzymes is that they produce tens of thousands of different natural products, many of which have industrial and biomedical applications (3), yet the substrates are limited to only a handful of allylic diphosphate compounds. Thus, for this superfamily, the challenge for functional assignment is determination of product specificity and not substrate specificity.

3. TARGET SELECTION

We used sequence similarity networks (SSNs) to visualize relationships among superfamily members (1,4). Figure 1a depicts a network of a representative set of 1771 sequences of the IS superfamily, which cluster into 5 distinctive subgroups that can be associated with the known catalytic reactions in each. Sequences with known structures and or functions (< 7% of superfamily members) were used to hypothesize initial isofunctional families within the subgroups, with each family and subgroup being defined by a group of sequences and a corresponding multiple sequence alignment and hidden Markov model (HMM). After initial families were created, hundreds of potential targets were suggested to cover the breadth of unknowns in the superfamily. Then, representatives for closely related sets were chosen based on the availability of DNA (to enable cloning), likelihood for successful protein production and structural characterization by crystallography or modeling, genetic tractability of species (for microbiology), criteria to enable *in silico* docking predictions of function, and experimental screening by enzymology (Figure 1b). The resulting set of 540 targets was submitted for analysis by EFI computational and experimental cores. We put special emphasis on selecting targets that catalyze the interesting cyclopropanation reactions (squalene/phytoene synthase like subgroup), cyclization reactions from the terpene synthase like 2 subgroup (Figure 1b), and a group of enzymes from the polyprenyl synthetase like subgroup that catalyze the transfer of prenyl groups to tryptophan residues in peptides. As the structures and functions of these targets are experimentally determined/validated, the results will be used to improve function prediction using modeling, docking, and bioinformatics. Automated protocols incorporating multiple sequence alignments and HMMs defined for each family and subgroup will be used to further curate isofunctional families and to develop annotation transfer rules so that the experimental information can be extrapolated to unknowns. Automated protocols will also be used to add new data as it comes available. Public access to the data, along with interactive versions of networks and search and analysis tools are provided by our Structure-Function Linkage

Database (http://sfld.rbvi.ucsf.edu) (5) and the EFI website (http://www.enzymefunction.org).

4. CONCLUSIONS

Organizing sets of homologous sequences into superfamilies provides a context for selecting targets for which experimental structural and functional characterization can best be leveraged to inform structure/function prediction for unknowns. Following a multidisciplinary approach that combines computation and experiment, the EFI aids in development of deep insight into the biochemical, metabolic and evolutionary aspects about enzyme superfamilies. Using the IS superfamily as an example, we show that this approach improves our success in predicting function for unknown enzymes discovered in the genome projects and, using automated update protocols, scales as the size of sequence databases increases.

5. FIGURES

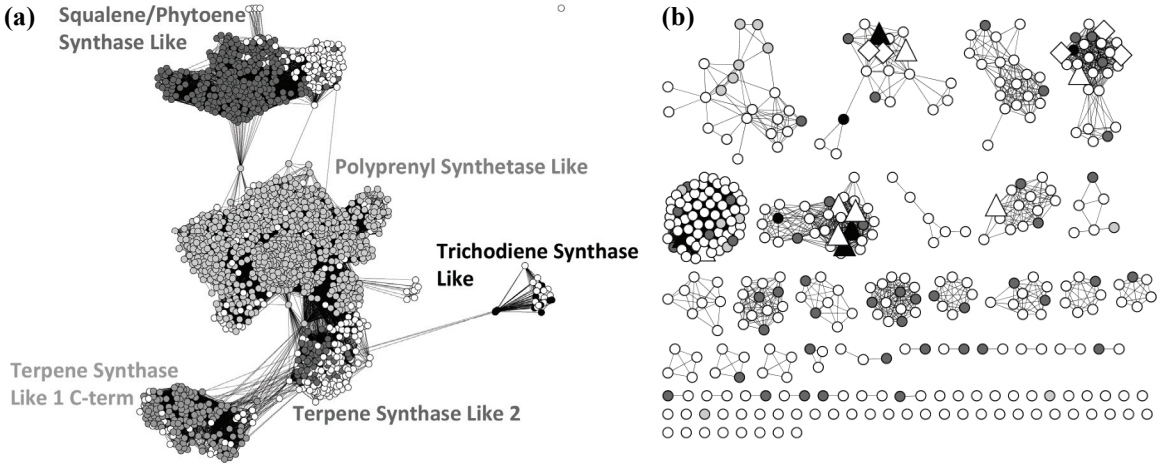


Figure 1. Sequence similarity networks (SSNs) for isoprenoid synthase superfamily members. Nodes correspond to sequences and edges to BLAST pairwise sequence relationships using scores better than a given threshold. (a) SSN generated from all-by-all pairwise BLAST comparisons of 1771 sequences from the IS superfamily filtered to 50% pairwise identity and thresholded at an E-value of $1x10^{-7}$. Sequences are colored according to the subgroups to which they belong as determined by matches to HMMs. White nodes represent superfamily members that do not match any of the subgroup HMMs. (b) SSN of all 374 sequences of the terpene synthase-like 2 subgroup thresholded at an E-value of $1x10^{-49}$. Triangles: known functions, diamonds: known structures and functions, black: past targets, light grey: current targets from genetically tractable organisms; dark grey: current targets for gene synthesis.

6. REFERENCES

1. Brown SD and Babbitt PC. 2012. Inference of functional properties from large-scale analysis of enzyme superfamilies. *Journal of Biological Chemistry* 287:35-42.

2. Gerlt JA, Allen KN, Almo SC, Armstrong RN, Babbitt PC, Cronan JE, Dunaway-Mariano D, Imker HJ, Jacobson MP, Minor W, Poulter CD, Raushel FM, Sali A, Shoichet BK and Sweedler JV. 2011. The Enzyme Function Initiative. *Biochemistry* 50:9950-9962.

3. Wendt KU and Schulz GE. 1998. Isoprenoid biosynthesis: manifold chemistry catalyzed by similar enzymes. *Structure* 6:127-133.

4. Atkinson HJ, Morris JH, Ferrin TE and Babbitt PC. 2009. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One* 4:e4345.

5. Pegg SC, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE and Babbitt PC. 2006. Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry* 45:2545-2555.

Can we go beyond sequence similarity to predict protein function ?

Paolo Fontana[1] , Tiziana Sanavia[2], Andrea Facchinetti[2], Enrico Lavezzo[3], Marco Falda[3], Duccio Cavalieri[1], Barbara Di Camillo[2], Stefano Toppo[3]*

[1] Istituto Agrario San Michele all'Adige Research and Innovation Centre, Foundation Edmund Mach, via E. Mach 1, I-38010, San Michele all'Adige (Trento), Italy
[2] Department of Information Engineering, University of Padova, via Gradenigo 6, I-35131, Padova, Italy
[3] Department of Molecular Medicine, University of Padova, v.le G. Colombo 3, I-35131, Padova, Italy

*To whom correspondence should be addressed: stefano.toppo@unipd.it

1. INTRODUCTION

Recent results of CAFA experiment have given us the unique opportunity to rethink the strategy used so far to make function prediction. Our approach, Argot2 (1,2), is mainly based on evaluating sequence similarities provided by BLAST and HMMER searches vs Uniprot (3) and PFAM (4) databases respectively. After extracting from GOA (5) database the Gene Ontology (GO) annotations of sequence hits and empirically weighting their e-value scores, we let Argot2 algorithm to make some considerations about how these annotations distribute and cluster in the graph by means of semantic similarity. Looking at the obtained results we have realized that, to some extent, PFAM searches have been used improperly. What is most surprising is that, rather than extending GO annotations, the PFAM results have simply reinforced those hits already found by BLAST. It is as if we have some GO terms weighted twice and consequently their final scores have been overestimated, while the improperly added terms were simply trimmed because belonging to poorly weighted paths. The take home message is that we are still bound to sequence similarity paradigm and we are still entrapped in this idea. CAFA has confirmed that top performing methods still rely on this. After all, what else can we do if the only thing we have is the amino acid sequence of the target to predict ? Can we go beyond sequence similarity to predict protein function?

Exploring weak signals of similarity is dangerous as false positive hits are the majority. On the other hand, we know that function, as well as the protein fold, can be conserved despite a great divergence in amino acid sequence. For these reasons we are figuring out how this peculiarity can be exploited. We are planning to look the other way round i.e. how function is distributed in GOA using the semantic similarity, in order to investigate if there is a correlation with sequence similarity or exceptions to take care of.

We are exploring the distribution of GO terms in GOA database and the possible correlation of their homogeneity with their sequence similarity (see Fig. 1). Eukaryotic proteins of same length have been extracted from Uniprot and clustered using CD-HIT (6) at 100%, 80%, 60%, 40%, and 20% sequence identity (Uniprot_100, Uniprot_80, Uniprot_60, Uniprot_40, Uniprot_20). Real median of sequence identity of the obtained groups has been recalculated and reported in the x-axis of Fig 1. In order to understand how domains and protein architectures are built looking at their associated functions, the same analysis has been performed for PFAM and TIGRFAMs (7) (the “equivalogs” groups) databases. The sequence identity has been calculated and reported in the x-axis of Fig.1 for TIGRFAMs, PFAM-A seed models (PFAM seed), and Protein architectures extracted from PFAM-A models (i.e. proteins having the same domains in the same order and number - PFAM Arch.). For each group of proteins the level of functional homogeneity has been assessed using the semantic similarity based on Lin's formula (8) and the percentage of these "homogeneous" groups has been reported in the y-axis of Fig. 1.

The present scenario of how GO terms are spread in the protein databanks seems to demonstrate that function is conserved up to 40%-50% sequence identity but dramatically drops when moving to 30% or lower values. Indeed, this result may be biased given that automatic annotations of IEA terms in GOA are mainly based on sequence similarity. So, if on the one hand it is not surprising to see that function conservation drops at low levels of identities, on the other hand it is interesting to observe that the majority of groups of proteins are still semantically homogeneous.

The final outcome, though preliminary, will help us to design a better solution in the future Argot3 algorithm hoping to have a more comprehensive view to automate functional inference even for those difficult cases that do not share high sequence similarities with known proteins.

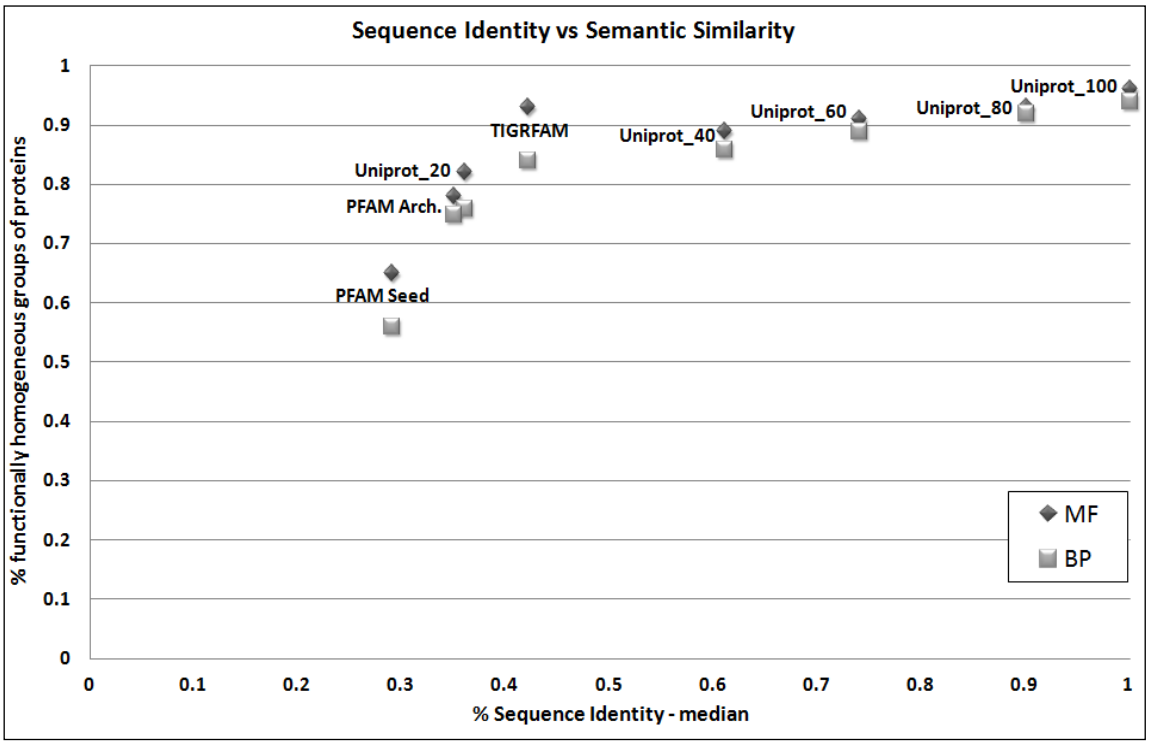


Fig. 1: percentage of sequence identity vs percentage of semantically homogeneous functions calculated for groups of proteins in PFAM, TIGRFAMs, and Uniprot databases. The data are reported for both Molecular Function (MF) and Biological Process (BP). See text for details

2. REFERENCES

1. Falda M., S. Toppo, A. Pescarolo, E. Lavezzo, B. Di Camillo, A. Facchinetti, A. Cilia, R. Velasco, and P. Fontana, *Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms.* Bmc Bioinformatics, 2012. **13**(4).
2. Fontana P., A. Cestaro, R. Velasco, E. Formentin, and S. Toppo, *Rapid Annotation of Anonymous Sequences from Genome Projects Using Semantic Similarities and a Weighting Scheme in Gene Ontology*. Plos One, 2009. **4**(2).
3. Bairoch A., R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, et al., *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2005. **33**(Database issue): p. D154-9.
4. Punta M., P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, et al., *The Pfam protein families database.* Nucleic Acids Research, 2012. **40**(D1): p. D290-D301.
5. Dimmer E.C., R.P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M.J. Martin, B. Bely, P. Browne, W. Mun Chan, R. Eberhardt, et al., *The UniProt-GO Annotation database in 2011.* Nucleic Acids Res, 2012. **40**(Database issue): p. D565-70.
6. Huang Y., B. Niu, Y. Gao, L. Fu, and W. Li, *CD-HIT Suite: a web server for clustering and comparing biological sequences.* Bioinformatics, 2010. **26**(5): p. 680-2.
7. Haft D.H., J.D. Selengut, and O. White, *The TIGRFAMs database of protein families.* Nucleic Acids Res, 2003. **31**(1): p. 371-3.
8. Lin D., *An Information-Theoretic Definition of Similarity*, in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML'98)*. 1998, Morgan Kaufmann Publishers Inc. p. 296-304.

Computational Function Predictions for Moonlighting Proteins

Ishita Khan 1, Meghana Chitale 1, Catherine Rayon 3, & Daisuke Kihara 2,1,*
1 Department of Computer Science, 2 Department of Biological Sciences, Purdue University, West Lafayette, IN, 47907, USA
3 Biologie des Plantes et Innovation, Université de Picardie Jules Verne, 80039 Amiens, France
*To whom correspondence should be addressed: dkihara@purdue.edu

1. INTRODUCTION

The importance of automated function prediction (AFP) methods is increasing because of the rapid accumulation of genome sequencing data. The overwhelming developments of genome sequencing technologies over the last few years have boosted the development of computational techniques and resources for protein function prediction. The traditional sequenced-based AFP methods are based either on the concept of homology or motif/domain matches. Some recent AFP methods use the hierarchy of the Gene Ontology (GO) or the phylogenetic trees. There are other function prediction methods that consider other types of data, such as gene co-expression patterns, protein-protein interaction networks, or 3D structures of proteins (1).

Although existing AFP methods have shown numerous successful predictions, moonlighting proteins may pose a challenge as they are known to show more than one function that are diverse in nature (2-3). The varied functional behavior of these proteins can be due to a difference in sub-cellular localization, expression by different cell types, binding of a cofactor, oligomerization, complex formation, or multiple binding sites. Moonlighting proteins have found to be involved in molecular functions ranging from diseases and disorders to immune systems.

To lay the framework towards computational function predictions for moonlighting proteins, in this work we have collected moonlighting proteins with experimental evidences from literature and identified GO term annotations for them from the UniProt database. Using the dataset, we have 1) analyzed how diverse the moonlighting functions are; and 2) analyzed the ability of existing function prediction methods to provide correct diverse functions (4). We have benchmarked PSI-BLAST and two sequence based AFP methods developed in our group, the Protein Function Prediction (PFP) (5) and the Extended Similarity Group (ESG) method (6) on the function prediction accuracy of the moonlighting proteins.

2. RESULTS

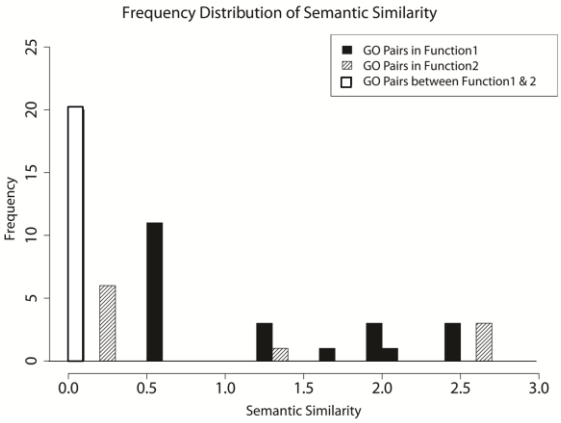
First, for all of the moonlighting proteins in the dataset, we computed the semantic similarity score (7) of pairs of GO terms that belong to the major moonlighting function (Function 1; F1), pairs that belong to the second function (F2), and pairs between F1 and F2. The semantic similarity score ranges from 0 to around 4.0, with a higher value indicating a match of the two terms at more detailed level of the GO hierarchy.

Figure 1 shows an example. The protein used here is mitochondrial aconitate hydrotase (UniProt ID: Q99798), whose major function (F1) is an enzyme in the tricarboxylic acid (TCA) cycle and the other function (F2) is involvement in the iron homeostasis. There are twelve GO terms that belong to F1 and five terms for F2. It is shown in that the pairs either in F1 or F2 have a higher semantic similarity score than GO pairs between F1 and F2. All the GO terms pairs between F1 and F2 have scores of 0, while the scores for the F1 pairs and the F2 pairs distribute in a higher score range, from 0.54 to 2.47 for F1 pairs and from 0.25 to 2.67 for F2 pairs.

Next, we benchmarked the performances of PFP, ESG, and PSI-BLAST in predicting the functional diversity of the moonlighting proteins. The 19 moonlighting proteins used here were taken from a review article (3). In the PSI-BLAST search, GO terms from top hits up to E-value of 0.01 were taken for annotating a target sequence. A GO term was ranked according to the score defined as –log(E-value) + 2 using the E-value of the sequence hit that has the GO term. PFP considers even very weak sequence hits (up to an E-value of 100), from which a score defined by –log(E-value) is assigned to the GO terms that belong to the sequence. For each GO term, its score is accumulated from all the sequences hits that have the GO term annotation. PFP also considers associations of GO term pairs observed in annotations in UniProt, which is used to give a partial score to highly associated GO terms to those directly annotating sequence

hits. ESG method, in contrast, performs iterative searches of PSI-BLAST and takes consensus GO terms from sequence hits. Thus, PFP is aimed for achieving a larger coverage of annotations while ESG achieves better precision.

In Figure 2, we plotted the recall for the three methods for each of the 19 moonlighting proteins separately. PFP showed higher recall than PSI-BLAST for almost all the cases (except for proteins 2 and 4, which are ties). ESG has a higher recall than PSI-BLAST for proteins 6, 12, 15, and 19, similar recall of predictions as PSI–BLAST for protein 17, and a lower recall than PFP and PSI-BLAST for the rest of the proteins.



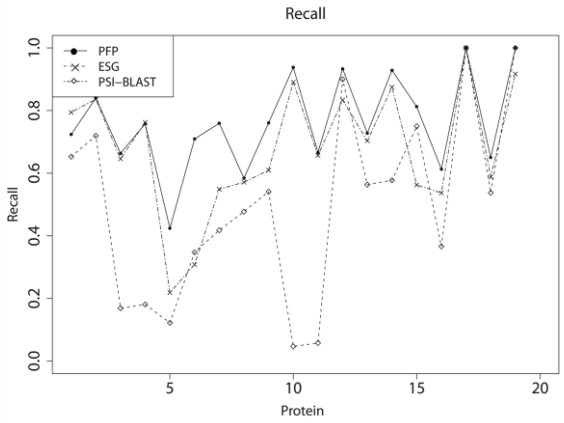


Figure 1 (left). Distribution of the functional similarity of GO terms within F1, F2, and between F1 & F2. Black bars, GO pairs in F1; gray, pairs in F2; white, pairs between F1 & F2.
Figure 2 (right). Recall of the three methods for only moonlighting functions (F1 and F2). Solid circle, PFP; cross, ESG; diamond, PSI-BLAST.

These results indicate that PFP can find moonlighting GO terms that are missed by regular PSI-BLAST searches for many cases. The strength of PFP is its coverage of a large number of sequences, by including weakly similar sequences into consideration for annotation transfer. On the other hand, ESG puts more weight on the consensus GO terms that are encountered in multiple iterations. Therefore, ESG fails to detect the functional variations in a number of cases. These results suggest that the functional diversity of the moonlighting proteins could be captured by considering a broad range of weakly similar sequences.

3. REFERENCES

1. Hawkins T. and Kihara D. 2007. Function prediction of uncharacterized proteins. *J Bioinformatics and Computational Biology* 5: 1-30.

2. Jeffrey C.J. 2004. Moonlighting proteins. *Trends in Biochemical Sciences* 24: 8-11.

3. Huberts D.H and van der Klei I.J. 2010. Moonlighting proteins: an intriguing mode of multitasking. *Biochim Biophys Acta* 1803: 520-525.

4. Khan I., Chitale M., Rayon C., and Kihara D. 2012. Evaluation of function predictions by PFP, ESG, and PSI-BLAST. *Submitted.*

5. Hawkins T., Chitale M., Luban S., and Kihara D. 2009. PFP: Automated prediction of gene ontology functional annotations with confidence scores. *Proteins: Structure, Function, and Bioinformatics* 74: 566-582.

6. Chitale M., Hawkins T., Park C., and Kihara D. 2009. ESG: Extended similarity group method for automated protein function prediction. *Bioinformatics* 25: 1739-1745.

7. Schlicker A., Domingues F.S., Rahnenfuhrer J., and Lengauer T.A. 2006. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* 7:302.

Lessons from the first critical assessment of functional annotation

Jesse Gillis, Paul Pavlidis*

Centre for High-Throughput Biology and Department of Psychiatry, University of British Columbia, 177 Michael Smith Laboratories 2185 East Mall, University of British Columbia, Vancouver, V6T1Z4, Canada

*To whom correspondence should be addressed: paul@chibi.ubc.ca

1. BACKGROUND

The assignment of gene function remains a difficult but important task in computational biology. The establishment of the first Critical Assessment of Functional Annotation (CAFA) was aimed at increasing progress in the field. We present an independent analysis of a portion of the results of CAFA, aimed at identifying challenges in assessment and at understanding trends in prediction performance. We found that well-accepted methods based on sequence similarity (i.e., BLAST) have a dominant effect. Many of the most informative predictions turned out to be either recovering existing knowledge about sequence similarity or were "post-dictions" already documented in the literature. These results indicate that deep challenges remain in even defining the task of function assignment, with a particular difficulty posed by the problem of defining function in a way that is not dependent on either flawed gold standards or the input data itself. In particular, we suggest that using the Gene Ontology (or other similar systematizations of function) as a gold standard is unlikely to be the way forward.

CAFA provided a unique opportunity to evaluate computational gene function assignments. Our results provide some new insights into the behaviour of gene function prediction methods, and into the challenges in providing an adequate and fair evaluation. We focus to some extent on comparing CAFA to CASP, and where helpful lessons could be learned, as categorized in the following sections.

2. TASK CATEGORIZATION

One area where CAFA could follow CASP is in the definition of tasks. Currently CASP differentiates between three categories of tasks, all of which have direct analogies with function prediction tasks. The CASP "template-based" prediction task is analogous to the case of trying to predict function when the gene has sequence similarity to already functionally annotated genes. In such cases, methods like BLAST provide a baseline for what can be learned readily. Our analysis shows that many of the CAFA targets already had "IEA" functions assigned, and to an extent CAFA successes are simply recovering these. Thus perhaps unsurprisingly, BLAST did well in the part of CAFA we had access to, and we expect that other high-scoring methods are using sequence similarity information. Tasks which exploit sequence similarity should be considered a distinct category of function prediction problems. Similarly, the CASP "template-free" prediction task is akin to the task of predicting gene function when no sequence similarity information is available (or at least, not used).

The CASP "structure refinement" task might be analogous to the task of "function refinement" where an already functionally annotated gene is given new or more specific functions. We believe this could be treated as is a different task from assigning functions to a completely unannotated "orphan" gene (not having even IEA annotations). Among methods that fall into this category are those which use GO itself as a measure of "guilt". Thus if two genes share nine out of ten GO terms, the tenth one is a pretty good bet. Even if they don't explicitly rely on existing annotations, algorithms that are good at "refinement" might not be very good at "template-based" assignment (and vice versa).

We propose that some scheme like this be adopted for future CAFA assessments, to more clearly differentiate between cases where sequence similarity is highly informative and those where it is not, and possibly to extend the competition to include targets which already have some functions assigned with "strong" evidence codes.

3. THE IMPORTANCE OF EVALUATION METRICS

Over the years, CASP has modified its assessment metrics and now has an agreed-upon set of metrics. Our results show that the gene-centred performance metric initially proposed for CAFA is unsatisfactory. This is illustrated by the fact that by this measure, a null "prediction method" outperforms most methods. The problem with the CAFA score is that it is not comparative across genes. When one is predicting a function for a gene, the goal is to say that "this gene has function X more than other genes do" in some sense.

Otherwise, the definition of function becomes degenerate, and simply assigning all genes the same functions becomes reasonable.

We have proposed two alternative measures, one which is gene-centric and focuses on the information content of a prediction, and a standard metric (area under ROC curve: AUROC) which is function-centric. The information-based metric is implicitly comparative, because it uses information on the distribution of GO terms across genes as well as a threshold set by the null predictor. The AUROC metric also ranks genes against each other. By these measures, it can be seen that the prediction algorithms (including BLAST) are providing meaningful performance. The problem with the function-centric measure is that it depends on having more than one prediction for the function to be scored, which cannot be guaranteed given the nature of the CAFA task. The differences in annotation practices for different organisms (notably for E. coli in the current data) make assessment even harder, as criteria vary for what is considered good annotations.

4. THE POWER OF AGGREGATION

In recent years, the top algorithms for CASP have tended to be meta-algorithms which aggregate the results of individual methods. If our experience is representative, the same is likely to be true for CAFA. The aggregate algorithm from the CAFA submissions we evaluated outperforms all the individual algorithms. The reason for this is apparently because aggregation allows a few “confident” predictions to rise to the top, while less confident predictions (which turn out to be poor) are “averaged out”.

5. THE BENEFITS OF HAVING A CLEAR GOAL

The points raised thus far are predicated on the idea that function prediction is like protein structure prediction. However, in a fundamental way this is not the case, at least not yet. Algorithms that perform well in CASP are considered to do well at “structure prediction”. That is, the CASP tasks are well-aligned with what the field agrees the “real life” task is. This is basically because protein structure is fairly easy to define. In contrast, “gene function” does not have an agreed-upon definition. Certainly there is no consensus that the Gene Ontology is even close to biological reality, rather than just being convenient. Since function assignment/prediction methods always use experimental data as inputs, there may be more value in simply trusting those data than in trying to “align” predictions to a gold standard that is acknowledged by its creators to be problematic for such uses. Tuning algorithms to be good at predicting GO annotations is probably never going to be satisfying. It is worth mentioning that there are “function prediction” tasks that are not based on GO (or similar schemes) in the same way as CAFA. For example, some groups attempt to predict mutant phenotypes. The roles of the issues we raise in such situations are not entirely clear, but we note that the types of data used are the same as those used in the CAFA-style annotation task, and GO often figures prominently in such work, especially as a source of validation .

6. PREDICTING EVIDENCE CODES AND “POST-DICTION”

With the caveat that CAFA’s evaluation is based on a relatively small number of proteins, there are some important themes that emerged in terms of which informative predictions were made. The evidence strongly suggests that a major factor is the availability of sequence similarity information. Finding a set of proteins which are not annotated at all was difficult, so many of the evaluation targets already had “IEA” annotations (presumably often based on BLAST or a similar approach). The successful predictions are in part simply guessing that those existing annotations are likely to be supported by experimental evidence once they are tested, and thus upgraded in GO.

The fact that many of the most predictable annotations were based on literature reports that predate CAFA further suggests that a bottleneck in filling in GO is information retrieval from the literature, not prediction per se. Strictly speaking, many of the CAFA evaluation targets are “post-dictions”. The short time window available to CAFA probably helped ensure this would be a factor; there was little chance that many experimental reports would be published and also curated in a six month period. The organizers were aware of this, and it is unlikely that CAFA participants would have been able to efficiently exploit the existence of publications describing the functions of proteins in the target set. On the other hand, for all we know some of the entries may have used text mining methods as a tool for making predictions. This might be considered yet another category of automated annotation task. But we stress that all the predictions are based on experimental data of one type or another, so this distinction may not be helpful. This returns us to the issue of the relationship between function prediction and GO. If computational predictions are based on experimental data that could be used by curators to populate GO, then the task of prediction is reduced to simply copying that information into GO (with appropriate evidence codes), rather than considering GO to be some independent entity that algorithms should attempt to match.

Sources of Experimental Function Annotation in UniProt-GOA, Implications for Function Prediction

Alexandra M. Schnoes (1), Alexander Thorman (2), and Iddo Friedberg (2,3)
1 Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA
2 Department of Microbiology, Miami University, Oxford, OH USA
3 Department of Computer Science and Software Engineering, Miami University, Oxford, OH USA
*To whom correspondence should be addressed: alexandra.schnoes@ucsf.edu

1. INTRODUCTION

Computational protein function prediction programs rely upon well-annotated databases for testing and training their algorithms. These databases, in turn, rely upon the work of curators to capture experimental findings from the scientific literature and apply them to protein sequence data. However, due to high-throughput experimental assays, it is possible that a small number of experimental papers could dominate the functional protein annotations collected in databases. Here we investigate just how prevalent is the "few papers – many proteins" bias and examine the annotation of experimental protein function in the UniProt Gene Ontology Annotation project (GOA). We find that for several important model species, a significant fraction of the annotations available are provided by only a few dominant papers (Figure 1). Given that most high-throughput techniques can find only one (or a small group) of functions, it appears that some level of experimental protein function annotation bias is unavoidable. We discuss how this bias affects our view of the protein function universe, and consequently our ability to predict protein function. Knowing that this bias exists and understanding its extent is important for database curators, developers of function annotation programs, and anyone who uses protein function annotation data to plan experiments.

2. FIGURES

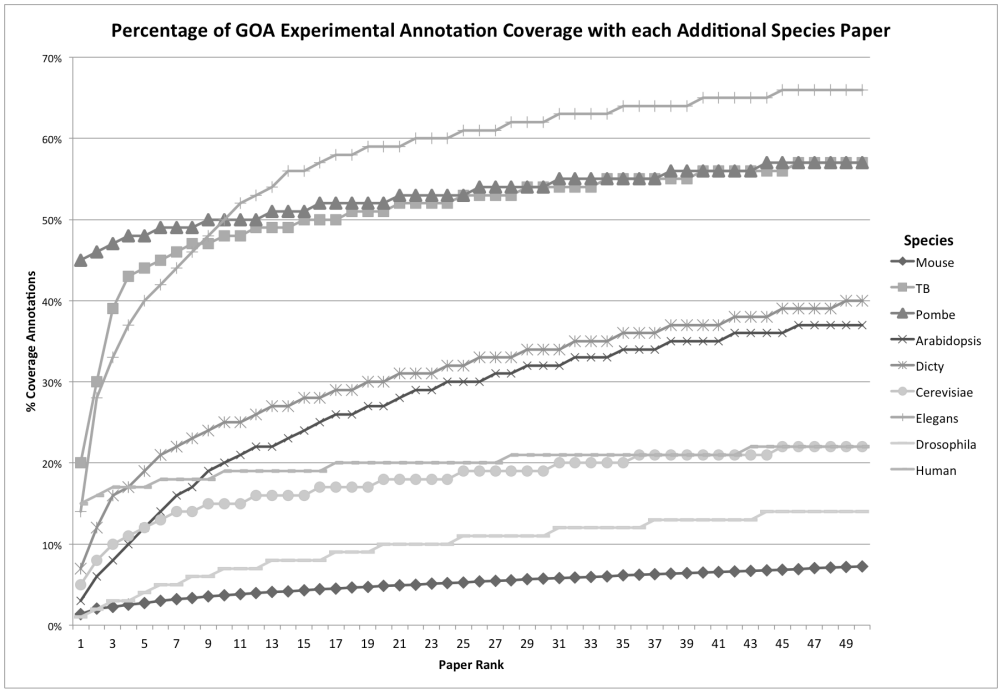


Figure 1: Percentage of GOA Experimental Annotation Coverage with each additional species paper. For

the nine species that populated the 50 papers in GOA with the most experimental annotations, an examination of annotation coverage from each annotation-producing paper was performed. The papers providing annotations in each species were ranked by how many experimental annotations in GOA they produced (i.e. each annotation had an evidence code from the following set: Inferred from Experiment, EXP; Inferred from Direct Assay, IDA; Inferred from Physical Interaction, IPI; Inferred from Mutant Phenotype, IMP; Inferred from Genetic Interaction, IGI; Inferred from Expression Profile, IEP). The paper that provided the most experimental annotations in each species was ranked as '1'. The percent coverage was then calculated for the top ranked paper and then recalculated with the addition of annotations from the next highest ranked paper and so on (percent annotation coverage per species = (number of annotations derived from a ranked paper + sum [number of all annotations produced from more highly ranked papers]) / total number of annotations for a species X 100%). For certain species, such as TB, a small number of papers have a large impact on the annotation coverage for that species. (Species abbreviations used: TB, *Mycobacterium tuberculosis;* Pombe, *Schizosaccharomyces pombe*; Arabidopsis, *Arabidopsis thaliana;* Dicty, *Dictyostelium discoideum*; Cerevisiae, *Saccharomyces cerevisiae* S288c; Elegans, C*aenorhabditis elegans;* Drosophila, *Drosophila melanogaster*.)

Information theory based metrics for the evaluation of GO term annotations

Wyatt Clark*, Predrag Radivojac
Indiana University, School of Informatics and Computing
To whom correspondence should be addressed: wtclark@indiana.edu

1. INTRODUCTION

While the Gene Ontology enforces a structure where child terms are always more specific than parent terms, one characteristic of the ontology is that terms at the same depth in the ontology exhibit a wide range of relative frequencies. Because of this, metrics which only consider the depth of a term in the ontology do not take into account the potential disconnect between the distance from the root of the ontology of a term and how informative it might be. Furthermore, instead of counting distinct objects as when used in information retrieval, metrics such as precision and recall do not take into account the lack of independence between parent and children terms in the ontology.

In order to account for the disparity between the depth of a term and how meaningful it is several metrics have been developed which take into account a term's information content [1, 3, 4, 2]. The information content of a term is inversely related to how commonly it occurs: rare terms contain more information whereas common terms are less informative. While information content metrics are successful in accounting for biases in the ontology due to uneven stratification of terms, due to the fact that they are formulated as a single value it is impossible to measure performance in a manner analogous to information retrieval concepts of precision, recall and specificity.

We have developed novel information theory based metrics for evaluating a set of GO annotations based on Kullback-Leibler divergence. Our metrics incorporate the advantages of semantic similarity but also have the added benefit of being interpretable as the information theoretic analogs of precision, recall and specificity. We compare and contrast our novel metric with some of the published metrics for evaluating GO annotations, pointing out advantages and disadvantages of each. Finally, we give a critique of evaluating a given metric based on its correlation with pairwise sequence similarity.

2. REFERENCES

[1] J.J. Jiang and D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *Arxiv preprint cmp-lg/9709008*, 1997.

[2] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on Machine Learning*, volume 1, pages 296–304. San Francisco, 1998.

[3] P. Resnik. Wordnet and distributional analysis: A class-based approach to lexical discovery. In *AAAI workshop on statistically-based natural language processing techniques*, pages 56–64, 1992.

[4] A. Schlicker, F.S. Domingues, J. Rahnenführer, and T. Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC bioinformatics*, 7(1):302, 2006.

Predicting Tissue Specificity from Protein Sequence

Sivan Goren, Yanay Ofran*
The Goodman Faculty of Life Sciences, Bar Ilan University, Ramat Gan 52900, Israel
*To whom correspondence should be addressed: yanay@ofranlab.org

1. INTRODUCTION

An important aspect of protein function in vertebrates is tissue specificity (1).
Genes specific to a given tissue are regulated by common mechanisms and hence prediction of tissue specificity is based mostly on identifying DNA sequence motifs upstream to the gene. We hypothesize that tissue specificity may be reflected not only in regulatory DNA elements but also in the biophysical characteristics of the proteins.
Protein features have been used in previous studies to predict protein function (2), protein-protein interactions (3) and subcellular localization (SCL) (4-7).The success of most SCL prediction methods is based, in part, on the notion that each subcellular compartment may constitute a slightly different microenvironment, with its own pH, osmolarity and viscosity. Proteins that are specific to a given subcellular compartment evolved to function in the specific conditions of that compartment and hence are expected to adjust their physicochemical traits accordingly. This is the rationale behind the fact that SCL prediction methods incorporate features such as amino acid composition and other sequence derived features (8, 9). Similarly, we hypothesize that, at least in some cases, tissue specific proteins may possess some common features that are identifiable from sequence and make them stable in a specific tissue.
We used expression data from the gene atlas to identify tissue specific proteins (10). Using these data we analyzed sequence derived features for tissue specific proteins and for ubiquitously expressed ones. We also analyzed and compared proteins specific to different tissues.
We demonstrate that such protein-derived features distinguish between proteins that are tissue specific and proteins that are expressed ubiquitously. Furthermore, we show that sequence-derived features can help determine whether a protein is specific to a given tissue. Interestingly, the predictions we make based on protein-derived features are at least as good as predictions that are based on DNA regulatory motifs. Thus, features extractable from the protein sequence may be used to improve the prediction of tissue specificity and promote our understanding of the mechanism of tissue specificity and its influence on the proteins function.

2. FIGURES

We used the data of tissue specificity to check whether we can predict from sequence derived features if a protein is tissue specific or not. We reached a precision of over 80% using 10-fold cross validation. Figure 1shows the precision to recall curve (PRC) and the ROC curve using random forest (RF) compared to a random model.

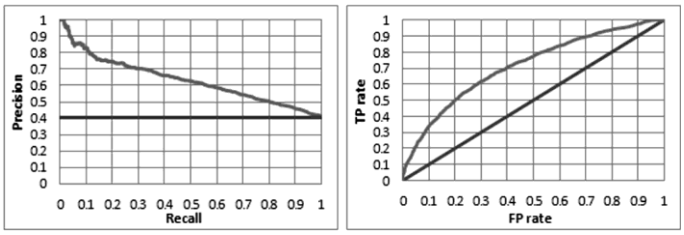


Figure 1: Classification of proteins as either "tissue specific" or "ubiquitous" based on the sequence derived features*. The darker bottom line represents the expected performance at random. The top gray line shows the performance of RF 10-fold cross validation. On the left is the PRC graph for the prediction of tissue specificity. On the right is the ROC curve.*

Our dataset is unbalanced, with more ubiquitous proteins than tissue specific ones, therefore the expected performance of a random assignment of proteins, would have a precision of ≈ 40%. We calculated precision and recall values for different raw output values of the RF. Starting at coverage of 85%, the precision is significantly better than random reaching a precision of 82% at 10% recall and an AUC of 71.24%.

Bone marrow, brain and blood were the three tissues for which we had the largest number of specific proteins. We extracted a list of 615 proteins specific to the bone marrow, 459 proteins specific to the brain, and 677 proteins specific to the blood. For each tissue we trained a RF model based on the same protein features. The results for each tissue are presented in Figure 2. For the brain we reach precision of 80% at 10% recall with an AUC of 65.11%. For bone marrow we reach precision of 60% at 10% recall with an AUC of 69.53%, and for blood we reach precision of 54% at 10% with an AUC of 65.74%. This performance is significantly better than random and is comparable to that of methods that are based on common regulatory motifs. Hence, sequence derived features could help predict tissue specificity.

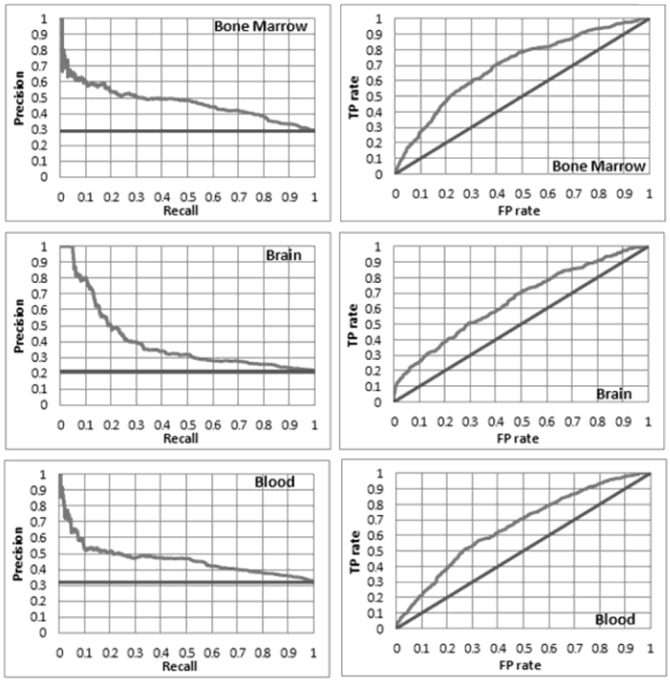


Figure 2: Results of RF models*; top – predicting bone marrow specific proteins; middle– predicting brain specific proteins; bottom – predicting blood specific proteins. As before, for all graphs the top line shows the test results of a 10-fold cross validation and the bottom line shows the expected random results*.

3. REFERENCES

1. Ponten, F., Gry, M., Fagerberg, L., Lundberg, E., Asplund, A., Berglund, L., Oksvold, P., Bjorling, E., Hober, S., Kampf, C., Navani, S., Nilsson, P., Ottosson, J., Persson, A., Wernerus, H., Wester, K., and Uhlen, M. (2009) A global view of protein expression in human cells, tissues, and organs. 5, 337.
2. Whisstock, J. C., and Lesk, A. M. (2003) Prediction of protein function from protein sequence and structure. 36, 307-340.
3. Bock, J. R., and Gough, D. A. (2001) Predicting protein--protein interactions from primary structure. 17, 455-460.
4. Cui, Q., Jiang, T., Liu, B., and Ma, S. (2004) Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. 5, 66.
5. Gardy, J. L., and Brinkman, F. S. (2006) Methods for predicting bacterial protein subcellular localization. 4, 741-751.
6. Guo, J., Lin, Y., and Liu, X. (2006) GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins. 6, 5099-5105.
7. Nair, R., and Rost, B. (2005) Mimicking cellular sorting improves prediction of subcellular localization.

348, 85-100.
8. Hoglund, A., Donnes, P., Blum, T., Adolph, H. W., and Kohlbacher, O. (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. 22, 1158-1165.
9. Hua, S., and Sun, Z. (2001) Support vector machine approach for protein subcellular localization prediction. 17, 721-728.
10. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* 101, 6062-6067.

"dcGO": a Domain-Centric Gene Ontology Predictor for Functional Genomics

Hai Fang*, Julian Gough
Computer Science dept., University of Bristol, UK
*To whom correspondence should be addressed: hfang@cs.bris.ac.uk

1. BACKGROUND

Computational/manual annotations of protein functions are one of the first routes to making sense of a newly sequenced genome. Protein domain predictions form an essential part of this annotation process. This is due to the natural modularity of proteins with domains as structural, evolutionary and functional units. Sometimes two, three, or more adjacent domains (called supra-domains) are the operational unit responsible for a function, e.g. via a binding site at the interface. These supra-domains have contributed to functional diversification in higher organisms. Traditionally functional ontologies have been applied to individual proteins, rather than families of related domains and supra-domains. We expect however that to some extent functional signals can be carried by protein domains and supra-domains, and consequently used in function prediction and functional genomics.
Results:

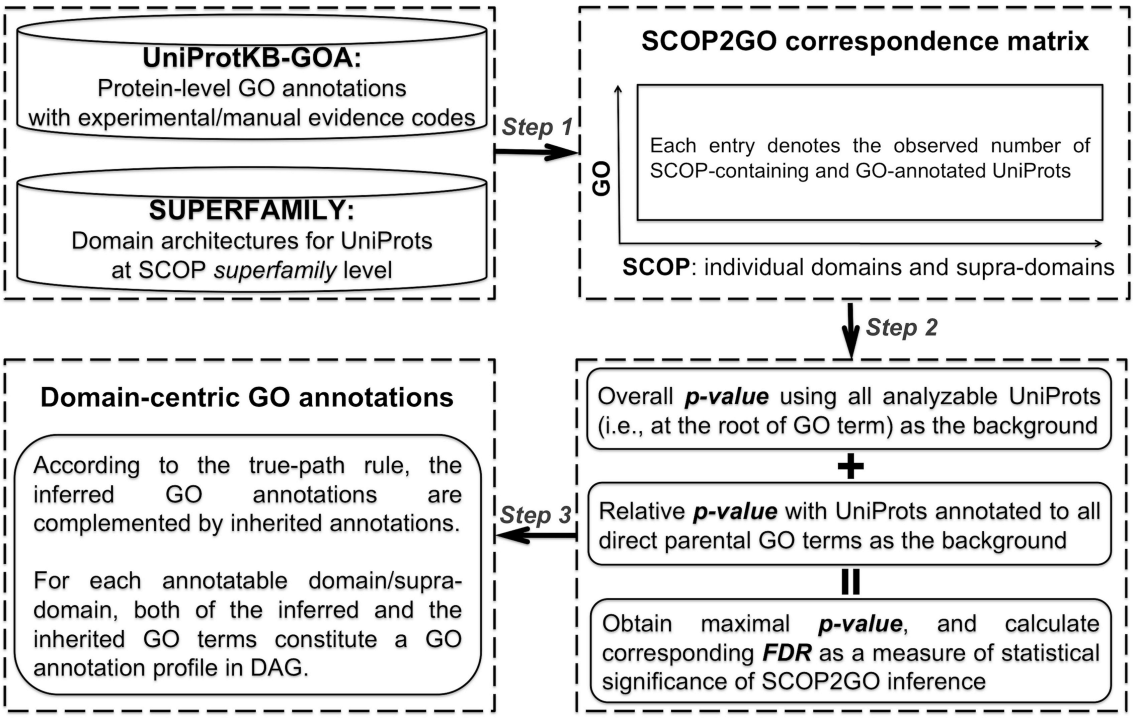
2. RESULTS

Here we present a domain-centric Gene Ontology (dcGO) perspective. We generalize a framework for automatically inferring ontological terms associated with domains and supra-domains from full-length sequence annotations. dcGO is available at http://supfam.org/SUPERFAMILY/dcGO.

This general framework has been applied specifically to primary Gene Ontology (GO) annotations from UniProtKB-GOA, and used to generate GO term associations with SCOP domains and supra-domains. The procedure is specifically tailored to the (directed acyclic graph) structure of the GO itself, and is designed to capture dcGO associations at the most relevant level. The resulting associations, via SUPERFAMILY, can be used to provide functional annotation to protein sequences. The functional annotation of sequences in the Critical Assessment of Function Annotation (CAFA) initiative has been used as a valuable opportunity us to validate our method and to be assessed by the community. The functional annotation of all completely sequenced genomes has demonstrated the potential for domain-centric GO enrichment analysis to yield functional insights into newly sequenced or yet-to-be-annotated genomes.

3. CONCLUSIONS

As functional units, domains offer a unique perspective on function prediction regardless of whether proteins are multi-domain or single-domain. The generalized framework we present has also been applied to other domain resources such as InterPro and Pfam, and other ontologies including eight phenotype and anatomy ontologies. The dcGO resource is routinely updated and holds great promise for a domain-centric functional understanding in the post-genomic era.



A flowchart illustrates a domain-centric GO approach to automatically infer GO annotations for individual domains and supra-domains.
This approach consists of three major steps, including (Step 1) the preparation of the correspondence matrix between domains/supra-domains and GO terms from protein-level annotations in UniProtKB-GOA and domain architectures in SUPERFAMILY database, (Step 2) two types of statistical inference followed by FDR calculation, and (Step 3) following the true-path rule to obtain the complete domain-centric GO annotations.

4. REFERENCES

1. de Lima Morais D.A., Fang H., Rackham O.J., Wilson D., Pethica R., Chothia C. and Gough J. 2011 SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res*, 39:D427-434.

2. Gough J, Karplus K, Hughey R and Chothia C. 2001 Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* , 313(4):903-919.

3. Automated Function Prediction: Critical Assessment of Function Annotations (CAFA). [http://biofunctionprediction.org].

Flexible Graphlet Kernels for Functional Residue Prediction in Protein Structures

Graph kernels for learning and inference on sparse graphs have been widely studied. However, the problem of designing robust kernel functions that can effectively compare graph neighborhoods in the presence of incomplete and/or noisy data remains less explored. Here we propose two novel graph-based kernel methods for predicting functional residues from protein 3D structure. These methods were designed to add flexibility in capturing similarities between local graph neighborhoods as a means of probabilistically annotating functional residues in protein structures. We report experiments on four residue-level function prediction datasets: identification of catalytic residues, identification of zinc-binding sites and DNA-binding sites, and phosphorylation site prediction. Our graphlet kernels performed as good as or better than established sequence and structure-based approaches. Additionally, we present evidence that the flexible kernels enable simple and principled ways of incorporating evolutionary conservation information into classification while efficiently capturing neighborhood similarities in protein structures.