Sequence- and Structure-Based Functional Annotation

BioChemistry 218: Computational Molecular Biology

Simar J. Singh
Winter 2009

Introduction

An important goal of computational molecular biology is the development of tools to predict the function of a novel or newly identified protein. With the explosive expansion of sequencing data, particularly from world-wide large scale genomics and proteomics initiatives, the number of uncharacterized sequences has exponentially increased within the last decade. As of January 2008, there were more than 600 completely sequenced genomes of cellular organisms, contributing to more than five million unique protein sequences within publicly accessible databases[1]. The experimental determination of the functions of all these proteins is both economically and logistically impractical. As such, only a fraction of sequences have had experimental confirmation of their function. In fact, at present only roughly 20%, 7%, 10% and 1% of annotated proteins in the Homo sapiens, Mus musculus, Drosophila melanogaster and Caenorhabditis elegans genomes, respectively, have been experimentally characterized and annotated[2]. Thus there is a significant need for computational approaches to help direct the functional annotation of the life's vast proteome.

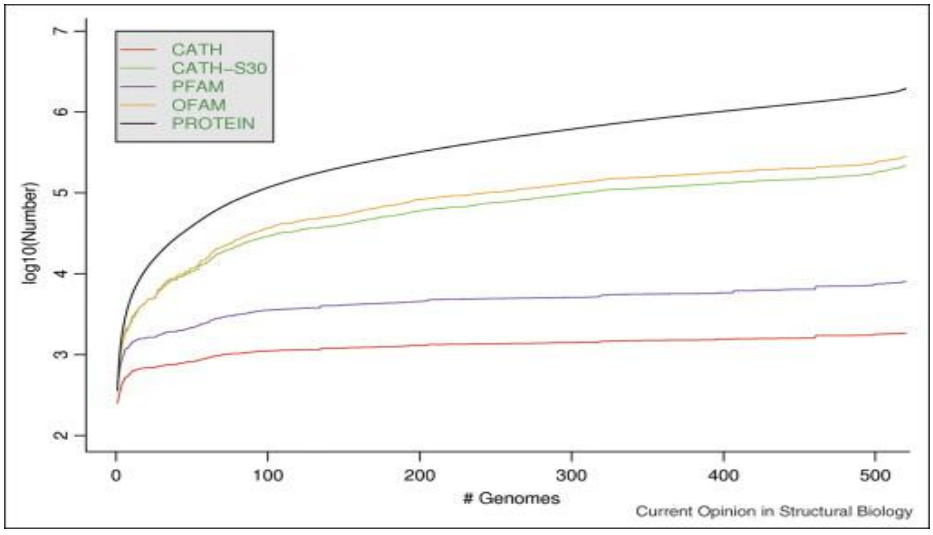

Figure 1 This figure shows the numbers (as a log scale) of genome sequences (PROTEIN), OFAMs (putative orthologous gene families), CATH-S30 (30% non-redundant sequence families), PFAM families, and CATH domain superfamilies, as more genomes are added to the database. (From Redfern & Orengo, 2008)[2]

Sequence based approaches

[1] Rison, S.C. et al. (2000) Comparison of functional annotation schemes for genomes. Funct. Integr. Genomics 1, 56–69

[2] O.C. Redfern *et al.*, Exploring the structure and function paradigm, *Curr. Opin. Struct. Biol.* **18** (2008), pp. 394–402. and references therein

The easiest way to infer the function of an uncharacterized protein is through sequence similarity with a well characterized homologue. However, many newly identified sequences have few if any well characterized homologues. In these instances, computational methods can provide a first estimate of protein function. Some of these computational tools rely upon similarity grouping, phylogenomics, sequence patterns, sequence clustering, and machine learning (ML).While similarities due to common ancestry can often be identified by alignment techniques, either pairwise or profile-based, similarities produced by common selective pressures are more subtle and are best identified using ML techniques such as artificial neural networks, support vector machines (SVMs) or hidden Markov models (HMMs) adapted to the topology and sequential structure of protein specific functional patterns. These functional patterns can be local, taking the shape of linear motifs or regions, or they can be developed for more global features such as amino acid composition or pair frequencies, or by combinations of local and global features.

Similarity group methods

The results of a BLAST (Basic Local Alignment Tool) query against a public database can provide a picture of the functional properties of related sequences. By looking at the quality and number of hits, one gains an idea of how large or diverse the family of an uncharacterized protein is. Moreover, by looking at the descriptions of the hits, one can see how well annotated the family of an uncharacterized protein is. Moreover, the dramatic increase in sequence information, the introduction of the Position Specific Iterated Blast (PSI-BLAST), and the establishment of the Gene Ontogney (GO) project have made similarity group methods a powerful tool for inferring protein function.

The GO system provides researchers with the ability to screen microarray data for GO similarity in differentially expressed transcripts. The sequences returned in a similarity search will frequently be enriched in multiple different GO terms. The more often a particular function appears, and the higher the quality of the hit scores, the more likely a particular protein will share this function. Because all parents of a given GO term can be scored, more general functional annotations for the query sequence can be supported by the occurrence of multiple more specific annotations in the list of hits.[3] The GOtcha software was one of the first programs to implement this GO-based tool.[4]

[3] M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat. Genet.* **25** (2000), pp. 25–29.

[4] D.M. Martin *et al.*, GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes, *BMC Bioinformatics* **5** (2004), p. 178.

PFP, the most recent implementation, uses three rounds of a PSI-BLAST search and a liberal threshold of sequence similarity.[5] As such it is able to include the annotations of more remote homologues in the scoring process. An additional source of improvement is the use of a 'functional association matrix,' which can support a certain annotation based on its statistical co-occurrence with other annotated sequences of high-quality.[6]

The phylogenic approach

Phylogenic approaches can improve the transfer of annotations by incorporating details of evolutionary relationships between protein families. This helps address the difference between orthologous and paralogous proteins. That is, the difference between relatives linked by speciation and those linked by gene duplication. The phylogenomic method is based on a standard work-through.[7] This includes the identification of all homologoues of the query sequence and their alignment, the building of a phylogenetic tree from the homologues, the reconciliation of the tree, and the transferring of function from orthologues.[8] Nearly three decades ago, Goodman and colleagues outlined the crucial step of reconciling the tree, which refers to the marking of all bifurcations in a tree as either the product of speciation or gene duplication.[9]

A recent implementation of this idea is SIFTER, which tries to transfer GO annotations from orthologes and inparalogues from within the same domain family of the Pfam database.[10] It accounts for different evolutionary speeds within a family by statistically modeling the evolution of molecular function.[11] SIFTER, like the other similarity group methods, exploits the GO annotation structure and further weights inherited annotations by the reliability (GO evidence code) of the source annotation.[12] Through thorough benchmarking and validation studies, the SIFTER analysis has been shown to yield highly specific and accurate annotations of protein

[5] T. Hawkins *et al.*, PFP: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data, *Proteins* **74** (2008), pp. 566–582.
[6] Ibid
[7] E.L. Sonnhammer and E.V. Koonin, Orthology, paralogy and proposed classification for paralog subtypes, *Trends Genet.* **18** (2002), pp. 619–620
[8] Ibid
[9] M. Goodman *et al.*, Fitting the gene lineage into its species lineage. A parsimony strategy illustrated by cladograms constructed from globin sequences, *Syst. Zool.* **28** (1979), pp. 132–163.
[10] B.E. Engelhardt *et al.*, Protein molecular function prediction by Bayesian phylogenomics, *PLOS Comput. Biol.* **1** (2005), p. e45
[11] Ibid
[12] Ibid

function.[13] More recently, the SIFTER methodology has been integrated into a meta-server for eukaryotic sequence prediction by the AFAWE tool.[14]

At present, similarity groups are ready for whole-genome application, meaning that they are relatively fast to compute and exhibit moderate precision levels. Phylogenetic trees on the other hand, are still significantly slower, but provide more precise inferences of protein function. Much current work is aimed at extended both of these methods. The 'Extended Similarity Group' scheme is promising to be both more sensitive and accurate than its predecessor PFP.[15] Additionally, Brenner and colleagues are working to implement a scalable version of SIFTER that is amenable to whole genomes.[16]

Pattern-Based Methods

An alternative approach to using whole sequence homologues from public databases is the use of conserved pattern sequences or motifs. These conserved pattern sequences can often indicate the function of the entire protein, simply based on a few signature residues. For instance, active site motifs can yield important information as to the catalytic functions of many enzymes. Pattern based methods can focus on different levels of functional specificity, and are reflective of the different sizes and complexity of the patterns used. These patterns can include protein domains as well. The 'Protein Feature Ontogeny' makes it possible to annotate all these features in a formalized manner.[17]

PROSITE pioneered the use of conserved patterns nearly 20 years ago, and has been steadily maintained and improved ever since.[18] It scans a query sequence against short, position-specific residue profiles that are characteristic of individual protein families. For example, these conserved residue motifs may be representative of the active site of a class of similar enzymes. PROSITE is complemented by ProRule, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acid residues.[19]

---

[13] B.E. Engelhardt *et al.*, Protein molecular function prediction by Bayesian phylogenomics, *PLOS Comput. Biol.* **1** (2005), and references therein.
[14] A. Jocker *et al.*, Protein function prediction and annotation in an integrated environment powered by web services (AFAWE), *Bioinformatics* **24** (2008), pp. 2393–2394.
[15] O.C. Redfern *et al.*, Exploring the structure and function paradigm, *Curr. Opin. Struct. Biol.* **18** (2008), pp. 394–402.
[16] O.C. Redfern *et al.*, Exploring the structure and function paradigm, *Curr. Opin. Struct. Biol.* **18** (2008), pp. 394–402.
[17] G.A. Reeves *et al.*, The Protein Feature Ontology: a tool for the unification of protein feature annotations, *Bioinformatics* **24** (2008), pp. 2767–2772.
[18] N. Hulo *et al.*, The 20 years of PROSITE, *Nucleic Acids Res.* **36** (2008), pp. D245–D249.
[19] Ibid

A widely used gateway to pattern-based functional annotation is InterPro, which integrates collected patterns from different levels into hierarchically arranged database entries. The InterProScan server is a powerful meta-tool which scans a query sequence against a ten core member database from which the output is presented in a non-redundant manner.

Among the most important of the InterPro domain profiles members are Pfam, SUPERFAMILY, PRODOM, SMART and, GENE3D. Of the protein profiles, PANTHER, PIRSF, and TIGRFAM's are the most important. PRODOM is able to automatically cluster evolutionary conserved sequence clusters based on reiterative PSI-BLAST searches of the UnitProtKB database.[20] All the others use hidden Markov Models (HMMs) generated by multiple sequence alignments to represent protein families.[21]

In terms of domain similarity methods, the manually generated Pfam-A and the computationally generated Pfam-B together provide the highest coverage of known sequence space among the InterPro members.[22] They are able to classify sequences in a large number of relatively small, functionally conserved families. SMART has a similar goal, but consists of a smaller, but completely manually curated set of families.[23]

SUPERFAMILY and Gene3D are structurally based classifications. SUPERFAMILY assigns sequences to the domain families defined by SCOP (Structural Classification of Proteins), while Gene3D assigns sequences to domain families from the CATH database.[24,25] While these database are much bigger than those used in Pfam, they are able to contain vary remote homologues that are often only detectable by structural conservation.[26] Structure based inference will be discussed later in this paper.

The PANTHER database attempts to delineate functional divergence within homologous protein families containing metazoan members.[27] Expert curation is then able to split the families into functionally conserved subfamilies annotated with GO molecular function and biological process terms.[28] TIGRFAMs focus on functional conservation as well, but their families contain 'equivalogs', which are sequences of conserved molecular function, regardless of their

[20] C. Bru *et al.*, The ProDom database of protein domain families: more emphasis on 3D, *Nucleic Acids Res.* **33** (2005), pp. D212–D215
[21] N. Hulo *et al.*, The 20 years of PROSITE, *Nucleic Acids Res.* **36** (2008), pp. D245–D249.
[22] R.D. Finn *et al.*, The Pfam protein families database, *Nucleic Acids Res.* **36** (2008), pp. D281–D288.
[23] I. Letunic *et al.*, SMART 5: domains in the context of genomes and networks, *Nucleic Acids Res.* **34** (2006), pp. D257–D260.
[24] D. Wilson *et al.*, The SUPERFAMILY database in 2007: families and functions, *Nucleic Acids Res.* **35** (2007), pp. D308–D313.
[25] C. Yeats *et al.*, Gene3D: comprehensive structural and functional annotation of genomes, *Nucleic Acids Res.* **36** (2008), pp. D414–D418.
[26] O.C. Redfern *et al.*, Exploring the structure and function paradigm, *Curr. Opin. Struct. Biol.* **18** (2008), pp. 394–402.
[27] H. Mi *et al.*, The PANTHER database of protein families, subfamilies, functions and pathways, *Nucleic Acids Res.* **33** (2005), pp. D284–D288.
[28] Ibid

evolutionary relationship (ie. either orthologs, paralogues, and even the products of horizontal gene transfer).[29] PIRSF uses 'homeomorphic' families of homologues, in which all members show full sequence similarity and a similar domain architecture, however not all of them share functional conservation.[30]

Profile based methods for predicting protein function follow a two step procedure. They first generate highly specific enzyme family profiles, and then assign unknown proteins to these families. Some methods incorporating this approach are the Cat-Fam method and its predecessors EFICAz and PRIAM.[31,32]

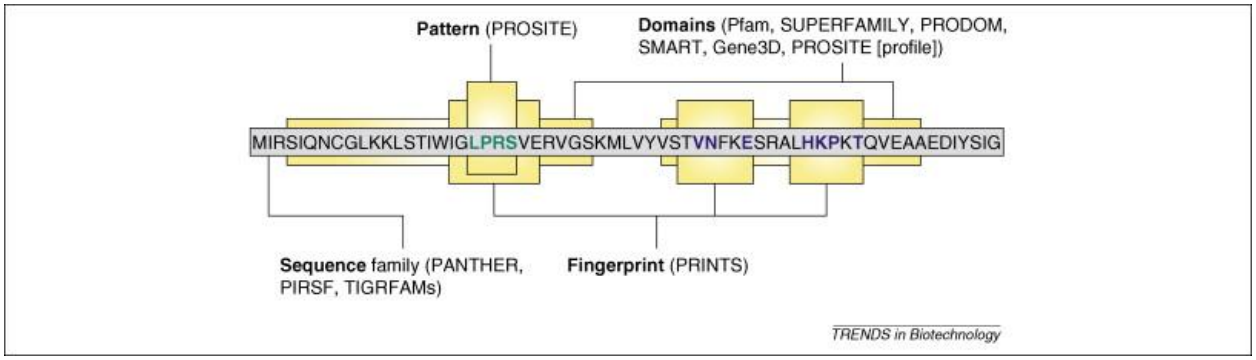

Figure 2 Conserved sequence patterns can be tied to protein function. This illustration shows an example sequence, with active site (green) and cofactor-binding site (blue) residues highlighted. The different InterPro member databases group protein sequences into families, based on conserved short patterns, fingerprints, domains or overall sequence similarity.

Clustering Approaches

A number of resources attempt to functionally annotate unknown proteins based on their clustering with characterized sequences. There are two principal approaches to sequence clustering. The first is clustering based on sequence similarity alone (homologues). The second is clustering based on supposed functional conservation (orthologues and paralogues). The first approach is incorporated in ProtoNet, which is an ambitious attempt at organizing sequences in trees of clusters.[33] ProtoNet uses a sophisticated clustering method to cut down on computational cost, and although it is completely automatic, it is in high agreement with manually curated

[29] D.H. Haft *et al.*, The TIGRFAMs database of protein families, *Nucleic Acids Res.* **31** (2003), pp. 371–373.
[30] C.H. Wu *et al.*, PIRSF: family classification system at the Protein Information Resource, *Nucleic Acids Res.* **32** (2004), pp. D112–D114
[31] C. Yu *et al.*, Genome-wide enzyme annotation with precision control: catalytic families (CatFam) databases, *Proteins* **74** (2008), pp. 449–460.
[32] A.K. Arakaki *et al.*, High precision multi-genome scale reannotation of enzyme function by EFICAz, *BMC Genomics* **7** (2006), p. 315
[33] N. Kaplan *et al.*, ProtoNet 4.0: a hierarchical classification of one million protein sequences, *Nucleic Acids Res.* **33** (2005), pp. D216–D218.

resources.[34] Moreover, annotation transfer based on these sorts of predominant features within known proteins, has been shown to be highly effective.[35] Similar to ProtoNet, CluSTr is tightly integrated with UniProt and IPI, thus providing the most complete picture of the sequence space.[36] However, in terms of basic benchmarking, ProtoNet produces the most accurate cluster.[37]

eggNOG, InParanoid, and OrthoMCL, are the three most widely used databases focusing entirely on clustering only orthologues and inparalogues.[38,39,40] Using orthologues and inparalogues as opposed to the entire sequence space is analgous to the difference between similarity and phylogenomic methods. eggNOG is a completely automatic way of clustering at different levels of taxonomy[38]. eggNOG assigns different domains to different orthologous groups, and as a result inherently accounts for sequence modularity. InParanoid is a repository of highly reliable pairs of orthologous proteins generated from model eukaryotic species. The MultiParanoid method expands these binary relationships to larger groups of orthologous proteins[39]. Both eggNOG and the InParanoid/MultiParanoid rely upon BLAST reciprocal best hits and/or triangular similarity relationships to nail down orthologues and inparalogues. OrthoMCL uses the popular Markov Clustering Alogorithm (MCL) to split coarse grained clusters[40]. A recent survey of similar clustering approaches found InParanoid and OrthoMCL to be the most accurate of these related resources[2].

Machine Learning

Support Vector Machines (SVMs) are predictors used to assign a given input (i.e. a sequence) to one of two classes (i.e. whether it has a particular function or not). The classification task is performed based on several features associated with each input (i.e. different physical/chemical properties). In Figure 3a, each circle represents a protein sequence from a training dataset, and the X and Y coordinates represent two distinct sequence features. The colors denote the presence (orange) or absence (green) of a certain function, for example a specific GO annotation. To train an SVM on this set of sequences, one seeks to determine the line that optimally separates the vectors of the functional from those of the non-functional

[34] Ibid
[35] O.C. Redfern *et al.*, Exploring the structure and function paradigm, *Curr. Opin. Struct. Biol.* **18** (2008), pp. 394–402.
[36] R. Petryszak *et al.*, The predictive power of the CluSTr database, *Bioinformatics* **21** (2005), pp. 3604–3609
[37] P.J. Kersey *et al.*, The International Protein Index: an integrated database for proteomics experiments, *Proteomics* **4** (2004), pp. 1985–1988.
[38] L.J. Jensen *et al.*, eggNOG: automated construction and annotation of orthologous groups of genes, *Nucleic Acids Res.* **36** (2008), pp. D250–D254.
[39] K.P. O'Brien *et al.*, Inparanoid: a comprehensive database of eukaryotic orthologs, *Nucleic Acids Res.* **33** (2005), pp. D476–D480
[40] F. Chen *et al.*, OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups, *Nucleic Acids Res.* **34** (2006), pp. D363–D368.

sequences, and maximizing the 'margin' (the arrow measuring the distance between the dashed lines) in the process. The size of the margin is constrained by the 'support vectors' which are the ones closest to the soild separation line (vectors on the dashed lines). After the training process, the SVM(s) can be used to characterize sequences that were not part of the training set, and the algorithm will vote either 'yes' or 'no' for a particular function.

The Neural Network (NN) approach is closely related to SVMs. Figure 3b shows a NN and its several layers of nodes. Here the input layer is orange, the output layers blue, and an arbitrary number of hidden layers are shown in green. Like real neurons, the nodes are connected. Each calculates an output from its different input values. The input nodes are fed with different signals (ie. sequence features shown as arrows). The output node 'votes' either yes or no (ie. for a certain GO term; this step is shown as a question mark). Networks are trained by adjusting the edge weights, which are the weights given to different input features (input layer) and their combinations (hidden layers). These weights are based on the known states of the output node for each training sequence. From this training an uncharacterized protein can have its functional features enumerated.

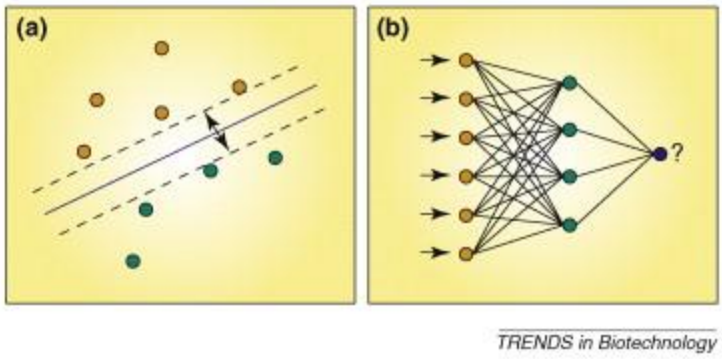

Figure 3 Machine learning paradigms. A) SVM and B) Neural Networks

The automated sequence annotation of sub-cellular localization is an important part of protein functional annotation. Perhaps the greatest success story in bioinformatics has been in the field of signal peptide prediction. Current algorithms are approaching reliability and accuracy levels of experimentally derived data[2]. Indeed, computational predictions of sub-cellular localization are often of higher quality than the underlying experimental data. The SignalP scheme was the first neural-network based approach that predicted the presence of a signal peptide and its cleavage site[2]. Another published machine learning approach that continue to perform well in this area is LOCTree, which is based on several binary SVM's that are arranged in three different decision trees specific for plants, non-plants, and prokaryotes.[41] Others include BaCeLo, which

[41] C.Z. Cai *et al.*, SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence, *Nucleic Acids Res.* **31** (2003), pp. 3692–3697.

is based on a decision tree of binary SVMs specific from animals, fungi, and plants, and TargetP, which is based on neural networks and has specificity similar to BaCeLo.[42]

Machine learning algorithms have also been useful for predicting membrane insertion. In general, topology predictors look for three important sequence characteristics of transmembrane alpha-helices. These are 1) a 20 amino acid long stretch of hydrophobic amino acids, 2) a flanking aromatic belt of tryptophan and tyrosine residues at the lipid-water interface, and 3) an overrepresentation of positively charged amino acids like lysine and arginine that are found in short cytoplasmic loops, referred to as the positive inside rule.[43] Incorporating machine learning algorithms and evolutionary history based on sequence profiles has increased the accuracy of predicting membrane protein structures to upwards of 80%.[44]

Newer algorithms are designed to pick up additional features that are characteristic of membrane function. These include lipid anchors and lipid modifications that are characteristic or particular membrane compartments.[45] Indeed machine learning has transformed the field of predicting subcellular localization from sequence identity. However, the power of these techniques is not limited to single functional sequence based prediction. Machine learning has the ability to integrate various functional signals, ranging from key catalytic residues to signals for sub-cellular localization and post-translational modifications in a more global inference of protein function.[46]

ORFan's are proteins or protein-encoding sequences within a genome that have no sequence similarity to proteins in other genomes. In the case that sequence queries return ORfan sequences, Machine-Learning approaches can provide useful hints regarding these unknown proteins' function. Such approaches try to learn highly specific, characteristic combinations of sequence features, or their intensities, that match functional assignments within a training set of known sequences. Support Vector Machines (SVM) or neural networks which support these specific classifiers are then used to assign functions to unclassified proteins in a probabilistic manner.[47]

The ProtFun server tries to accomplish this by assigning eukaryotic sequences to 1) one of 14 GO categories, 2) one of 12 cellular roles from the Riley scheme, and 3) an EC class if an enzyme. Moreover it can integrate multiple other tools that predict other protein properties, such

---

[42] Emanuelsson O: Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2007, 2:953-971.

[43] von Heijne G: Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* 1992, 225:487-494

[44] DT Jones: Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 2007, 23:538-544.

[45] Pierleoni A et al. PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* 2008, 9:392.

[46] Eisenhaber F: Post-translational modifications and sub-cellular localization signals: indicators of sequence regions without inherent 3D structure? *Curr Protein Pept Sci* 2007, **8:**197-203.

[47] L.J. Jensen *et al.*, Prediction of human protein function according to Gene Ontology categories, *Bioinformatics* **19** (2003), pp. 635–642

as hydrophobicity, post-translational modifications, sub-cellular localization, secondary structure composition, and potential transmembrane regions.[48] SVM-Prot is a similar predictor not limited to eukaryotic sequences and is able to specifically identify types of enzymes, receptors, kinases, transporters, and nucleic-acid binding proteins.[49]

ffPred is a PSI-BLAST dependent tool developed by Lobey and colleagues that attempts to make feature prediction.[50] It can consider disordered sequence regions and exploits the structure of the GO system to make accurate feature predictions.[51] ffPred is able to characterize eukaryotic sequences with an outstanding functional coverage. EzyPred is used to identify uncharacterized enzymes and integrates domain and evolutionary information for enzyme/non-enzyme characterization and a prediction of the first and second EC number classifications.[52] Moreover, EzyPred is able to accomplish these tasks with an accuracy of over 90%.[53]

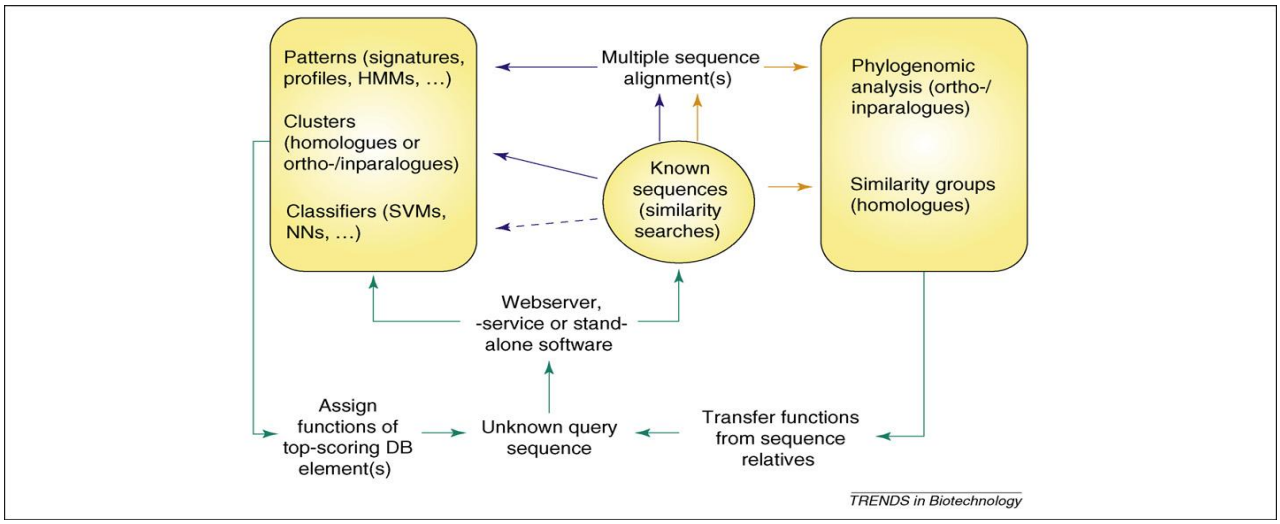

Figure 4. Integration of multiple sequence-based methods for function annotation (From Rentzsch & Orengo 2009)

Summary of Sequence-Based Protein Function Annotation

Clustering and similarity group methods only rely on one-to-one sequence comparisons, whereas pattern and phylogenomic approaches involve multiple sequence alignments. Machine Learning approaches rely instead on intrinsic sequence features but not sequence comparison to train their classifiers. In the case that the BLAST approach fails, one must work to find a consensus among these approaches to accurately infer protein function. Meta-servers that are

[48] Ibid
[49] C.Z. Cai *et al.*, SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence, *Nucleic Acids Res.* **31** (2003), pp. 3692–3697.
[50] A.E. Lobley *et al.*, FFPred: an integrated feature-based function prediction server for vertebrate proteomes, *Nucleic Acids Res.* **36** (2008), pp. W297–W302.
[51] Ibid
[52] H.B. Shen and K.C. Chou, EzyPred: a top-down approach for predicting enzyme functional classes and subclasses, *Biochem. Biophys. Res. Commun.* **364** (2007), pp. 53–59.
[53] Ibid

able to query multiple algorithms for sequence based protein prediction would be incredibly helpful in this task.

Various analyses have suggested that for functional transferability, a 40% pair-wise sequence identity can be used as a confident threshold to transfer the first three digits of an EC number, but to transfer all four digits of an EC number with at least 90% accuracy, over 60% sequence identity is needed[2]. Lower thresholds can be used (30% sequence identity) for domain relatives that share similar multidomain contexts[2]. Furthermore, because gene families evolve at different rates, family-specific thresholds are safer and lead to higher levels of functional annotation in many genomes (for example, a five-fold increase in GO annotations in D. melanogaster).[2]

Combining sequence- and structure- based prediction

Though functional assignment of uncharacterized proteins is commonly performed through sequence analysis, the assignment of function on the basis of homology can lead to incorrect or misleading annotations. Moreover, sequence-based predictions cannot identify new functions. The problem is compounded by the fact that there is no simple relationship between measures of sequence similarity, or sequence identity, and protein function. Highly similar proteins (with 60% sequence identity or greater) can catalyze distinct reactions, and highly divergent proteins can catalyze identical chemical reactions.[54]

Babbit and others have used the enolase superfamily to illustrate this point extensively.[55] In the enolase superfamily, a functionally diverse group of enzymes, misannotation of enzymatic function is severe. To address this, researchers have begun to incorporate structural and sequence information into their computational algorithms. Knowledge of the three-dimensional structure of a protein can provide crucial insight into its mode of action, but currently the structures of less than 1% of sequences have been experimentally solved, placing impetus on the development of computational processes that can guide our understanding of protein function.[56]

At present, structural models of most protein sequences are only available by homology modeling approaches. However, in principle, sufficiently accurate computational methods could enable the construction of models that could be used as surrogates for experimentally determined structures. This could be incredibly useful in, for example, drug discovery or for understanding sequence-structure-function relationships. Babbit and colleagues demonstrated that it was

[54] S.C. Pegg & P.C. Babbitt, Leveraging enzyme structure–function relationships for functional inference and experimental design: the structure–function linkage database, *Biochemistry* **45** (2006), pp. 2545–2555

[55] Ibid

[56] Grabowski M, Joachimiak A, Otwinowski Z, Minor W: Structural genomics: keeping up with expanding knowledge of the protein universe. *Curr Opin Struct Biol* 2007, **17:**347-353.

possible to predict the substrate specificity of a divergent member of the enolase superfamily based on docking against a homology model.[57] Subsequent enzymatic characterization and crystallographic analysis confirmed their predictions.[58]

The use of structural similarity in function prediction does face additional problems arising from artifacts of the crystallization procedure. A range of cognate and non-cognate ligands are used to stabilize the protein structure and facilitate the formation of crystals.[59] In some cases, any conformational change that occurs during substrate binding can cause significant changes to the overall structure.[60] Therefore, even structures of the same protein might exhibit significant structural differences when superimposed. However, structural data can be used to detect proteins with similar function whose sequences have diverged beyond the point where similarity that can be reliably detected using sequence comparison methods.[61] In general, approaches to predict function from structure rely on trying to find globally similar structures and then, if no match is found, to focus on any structural similarities between known or predicted functional sites.[62]

Predicting function by protein-fold comparison.

Several popular methods for aligning and quantifying structural similarity relationships along the entire length of amino acid sequences are: DALI, CE, SSAP, STRUCTAL and CATHEDRAL.[63] In these programs, attention is paid to both the quality of the superposition, and also the number of residues in the alignment. A key attribute of fold prediction is that the full alignment of two protein structures is not necessary for functional annotation. A recent approach scored the similarity of two proteins by simply comparing their internal residue contacts, which are those residues that co-locate within 8–10 Å in the structure.[64] This helps detect additional similarities over global alignment methods. This type of approach can improve the computing time of large multi-sequence analysis tremendously.

---

[57] S. Ojha, E.C. Meng and P.C. Babbitt, Evolution of function in the "two dinucleotide binding domains" flavoproteins, *PLoS Comput Biol* **3** (2007), p. e121.

[58] Ibid

[59] Polacco BJ, Babbitt PC: Automated discovery of 3D motifs for protein function annotation. *Bioinformatics* 2006, 22:723-730

[60] Ibid

[61] P.F. Gherardini and M. Helmer-Citterich, Structure-based function prediction: approaches and applications, *Brief. Funct. Genomic. Proteomic.* **7** (2008), pp. 291–302

[62] Ibid

[63] Redfern OC, Harrison A, Dallman T, Pearl FM, Orengo CA: CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput Biol* 2007, 3:e232

[64] Kolodny R, Koehl P, Levitt M: Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol* 2005, 346:1173-1188.

There does, however exist a problem with the use of protein folds to predict function. Namely, folds can have different functions. An analysis of the CATH database revealed that although most domains that share the same fold are associated with a single function, a small number of 'superfolds' (such as the Rossmann fold) can be associated with more than 50 different functions.[65] Furthermore, these superfolds are the most common folds and account for over 50% of domain sequences with predicted structures.[66] Moreover, in highly variable superfamilies such as the enolase family, different functions can evolve through secondary structure element insertion. Still, as a rule of thumb, most superfamilies with a high level of structural similarity also exhibit high functional similarity, but the supporting data is still at best sparse.

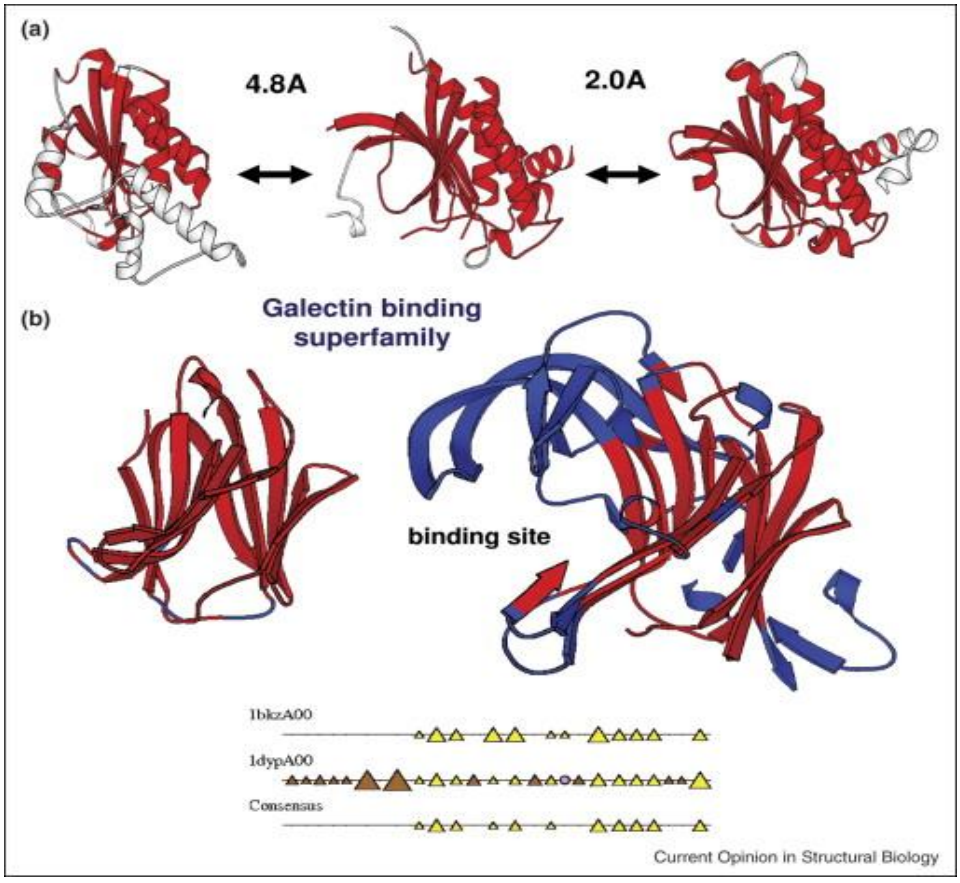

Figure 5 (a) Two domains which differ by less than 5 Å (SIMAX) can vary structurally and share the same fold. (b) Two domains from the galectin-type carbohydrate recognition domain superfamily.

Predicting function using local 3D templates.

[65] Redfern & Orengo CA: CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput Biol* 2007, 3:e232.

[66] L. Xie and P.E. Bourne, Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments, *Proc Natl Acad Sci U S A* **105** (2008), pp. 5441–5446

During evolution, the local microenvironment of a protein's active site must be preserved if that specific function is to be retained. This is so even if the folds or portions of the folds near the active site have been modified. Indeed, within an enzyme catalyst, a limited number of residues comprising the active site confer functional specificity. This introduces a problem when attempting to use whole fold comparison to assign function. Specifically, this approach is limited by the fact that small changes in a binding or active site can cause a divergence of function. As a consequence, there are several methods that focus on comparing smaller structural motifs associated with a specific function.

One such method is found within the Catalytic Site Atlas, in which up to six catalytic residues per enzyme can be manually annotated from the literature.[67] These annotations have been carefully transferred to close relatives using conservative PSI-BLAST profiles. The catalytic residues are used to construct structural templates, which are then queried by a fast search algorithm to compare unknown and known structures. Based on catalytic structure similarity, EC numbers can then be transferred.

This task is not as simple as it sounds. Catalytic residues are known to alter their geometry relative to each other upon substrate binding. This kind of information is difficult to predict without empirical data. Additionally, recognizing the correct relative can be difficult. The probability of matching small structural templates at random is high, raising the number of false positive hits during a query. One attempt to address this problem compares the local environments around known or predicted catalytic residues and the corresponding residues in the matched protein.[68] Using this localized approach, researchers can exploit the observation that the environment around the active site often exhibits higher sequence similarity than is evident across a global alignment of the query and matched structures. This provides a microenvironmental framework in which the catalytic residues must exist to confer a specific enzymatic function.

Other related methods make use of similar knowledge-based approaches to compare functional information (SITE records) that are found within the Protein Data Bank (PDB) structure files. However, this can be problematic because there are no set of standard rules for what information should be contained in the SITE records. Thus these files can contain a large amount of protein features (i.e. disulfide bridges, binding to unnatural ligands, information on mutated residues etc.) which may not be useful for functional comparison and annotation transfer.

Manually designing structural templates for a particular function is a time intensive process. As such, several projects are underway to develop novel algorithms to derive these structural

---

[67] C.T. Porter, G.J. Bartlett and J.M. Thornton, The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, *Nucleic Acids Res* **32** (2004), pp. D129–D133.

[68] Bagley SC, Altman RB: Characterizing the microenvironment surrounding protein sites. *Protein Sci* 1995, **4:**622-635

templates automatically. One approach seeks to detect common structural motifs through the pairwise comparison of side chain members within diverse members of protein families.[69] Then, these motifs can be scanned to see if they are present in uncharacterized proteins. Some algorithms also rely on common side-chain patterns within protein superfamilies, but make no assumptions about the nature or location of these motifs[69]. Since hydrophobic residues are usually found in a protein's core, they are usually excluded from template construction. The smaller the motif, the more difficult it is to distinguish between genuine similarities and false positives.

Babbit and colleagues have pioneered a novel approach of template building. Instead of using structurally conserved regions within protein families to identify 3D templates, they use random sequence-conserved residues in known enzyme structures to build their structural motif templates[69]. Interestingly, the best templates generally contain known functional residues, although there are also a few residues that have no known functional role. These non-reactive residues might afford a structural scaffold for catalytic or binding residues.

Comparing local structural features.

Comparing local structural features, such as the surface of a protein or pockets like the active site or ligand binding cleft, can yield important functional information regarding protein-protein interactions and small molecule binding. Enzymes create unique chemical environments within active sites and binding clefts that allow them to efficiently catalyze chemical reactions. Often these microenvironments are characteristic of a particular function and can be exploited to infer protein function.

In a fashion similar to 3D template searching, the binding sites of unannotated proteins can be compared against a library of known sites. This type of comparison is implemented in pvSOAR and related programs.[70] Some groups have expanded on this theme by including comparisons of the chemical properties of the amino acids in the binding site. Hydrophobicity and electrostatic charge are two conservative features that can be used to determine genuine functional homologues. Such a method is implemented by programs such as SiteEngine.[71] Similarly, binding sites with comparable physical/chemical properties can also be used to identify similar enzymatic functions. The eF-Site database builds on this by providing information of the

[69] Polacco BJ, Babbitt PC: Automated discovery of 3D motifs for protein function annotation. *Bioinformatics* 2006, 22:723-730.

[70] Binkowski TA, Freeman P, Liang J: pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucleic Acids Res* 2004, 32:W555-W558.

[71] Shulman-Peleg A, Nussinov R, Wolfson HJ: SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res* 2005, 33:W337-W341

electrostatic potential of active site surfaces that are then used to identify similar patterns of charge during binding and protein-protein interactions.[72]

Since these molecular interactions rely on electrostatic contacts between charged or polar residues, it is possible to use molecular cartography approaches to reduce protein surfaces to a spherical map.[73] By comparing the distributions of hydrophobic and charged residues within two maps, one can identify functional subgroups within protein families. These approaches also have application in the prediction of a protein's kinetic parameters. One can model protein electrostatics and predict the pKa values of ionizable groups within an active site by using theoretical microscopic titration curves.[74]

Figure 6 Work Through for structure based protein annotation. (From Babbit & Colleagues 2008)

[72] K. Kinoshita and H. Nakamura, eF-site and PDBjViewer: database and viewer for protein functional sites, *Bioinformatics* **20** (2004), pp. 1329–1330.
[73] J.M. Sasin, A. Godzik and J.M. Bujnicki, SURF'S UP! — protein classification by surface comparisons, *J Biosci* **32** (2007), pp. 97–100.
[74] Ibid

Servers for function prediction.

A number of servers have begun to incorporate structural properties in their function inference. Two of the most convenient servers include ProFunc[75] and ProKnow,[76] which extract both structural and sequence data during a query using several of the methods outlined above. ProFunc combines BLAST and HMM searches with 3D template and surface-cleft analysis.[75] ProKnow extends this approach by providing a probability model for GO annotations.[76] These and future meta-Servers have the potential to provide more accurate and complete functional annotations by combining several sequence and structure based methods for predicting protein function.

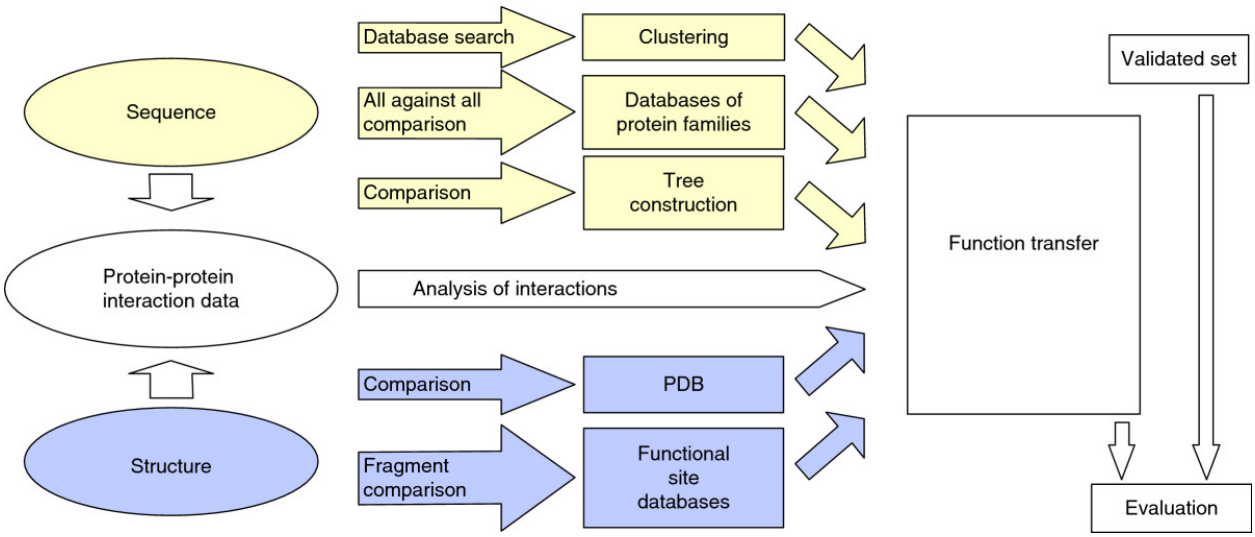

Figure 7 Integration of Sequence- and Structure-Based Methods of Protein Function Annotation (From Rentzsch & Orengo 2009)

Conclusion

With the advent of the structural genomics initiatives, an increasing number of protein structures are being experimentally determined while their function is still unknown. In these cases, function can sometimes be predicted by using the structure rather than the sequence of the protein. Analogous to sequence comparison, global comparisons can be made using fold-comparison methods, usually by identifying the individual structural domains in a protein, and local comparisons can be made using structural templates from the active site of enzymes. Other features that can be used for function prediction when a structure is available include conserved surface patches, clefts and electrostatic potential. Although 'inference through homology' is still

[75] Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res. 2005;33:W89–W93

[76] D. Pal and D. Eisenberg, Inference of protein function from protein structure, *Structure* **13** (2005), pp. 121–130.

the hallmark of function transfer, de novo methods of determining function based on machine learning schemes are gaining prominence.

The success of computation molecular biology as a field is due in part to its ability to integrate previously separate data sets in biologically meaningful ways. The combination of sequence- and structure-based function transfer approaches is a promising field for future research. This is due to their complementary nature. Sequence and structure similarity can provide a safe basis for function transfer and predict molecular interactions that hint at the biological pathways and processes in which an uncharacterized protein participates. A future goal will be to leverage computational methods to generate protein interaction datasets for metagenomic sequences. We may get to a point where these 'interactomes' may help us learn about evolution on a larger more systemic level, one ranging from the molecular to the organismal. Future methods of modeling interactions will have to overcome the fact that orthologues from different species have very little overlap in their substrates. This could pose problems much like those that plagued function transfer based on sequence similarity. Thus it is imperative that a multitude of both sequence and structural methods be used to find a consensus that provides the clearest picture of protein-protein interactions.

Though function annotation transfer has steadily improved over the years, there is still a critical need for standardization in the terms of annotation. While the GO system provides an important step in this direction, it is by no means perfect. Functional granularity still varies considerably at different branches of ontogeny. This makes the GO directed acyclic graph (DAG), which seeks to measure the functional similarity between two GO annotated sequences, largely unreliable. The field would benefit if the GO system settled on one universally supported functional distance measure. Moreover the GO consortium could actively support a benchmarking standard for function prediction. The field should embrace third-party benchmarking and validation studies which could help improve the accuracy of function transfer. Going one step further, GO efforts should be integrated with text mining and develop a standard set of publication rules as a basis for both manual curation and automatic inference. This would help the field keep pace with the increasing amount of newly generated sequence data.

For sequences that have already been produced, a two-pronged approach should be utilized. This should include both sequence and structural data. This would provide important guides for future experimental work on these uncharacterized proteins. For proteins whose functions have already been predicted, newer structural and sequence-based methods, like those outlined above could help repair misannotations in the public databases. In this manner a large number of proteins could be reannotated in a more accurate fashion.

Finally, it is imperative that these methods be publicized to the larger biological community. More work should be done to educate experimental biologists in the effective use of computational methods. The expansion of user-friendly meta-servers will help make this a

reality. This would greatly improve the applicability of new computation methods and provide the research community with powerful tools for understanding biology in a comprehensive, multidisciplinary manner.