Computational and Structural Biotechnology Journal 13 (2015) 182–191

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Biochemical functional predictions for protein structures of unknown or uncertain function

Caitlyn L. Mills, Penny J. Beuning, Mary Jo Ondrechen *

Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02115, United States

ARTICLE INFO

Article history:
Received 3 December 2014
Received in revised form 6 February 2015
Accepted 11 February 2015
Available online 18 February 2015

Keywords:
Structural genomics
Protein function prediction
Local structure methods
Computational chemistry

ABSTRACT

With the exponential growth in the determination of protein sequences and structures via genome sequencing and structural genomics efforts, there is a growing need for reliable computational methods to determine the biochemical function of these proteins. This paper reviews the efforts to address the challenge of annotating the function at the molecular level of uncharacterized proteins. While sequence- and three-dimensional-structure-based methods for protein function prediction have been reviewed previously, the recent trends in local structure-based methods have received less attention. These local structure-based methods are the primary focus of this review. Computational methods have been developed to predict the residues important for catalysis and the local spatial arrangements of these residues can be used to identify protein function. In addition, the combination of different types of methods can help obtain more information and better predictions of function for proteins of unknown function. Global initiatives, including the Enzyme Function Initiative (EFI), COMputational BRidges to EXperiments (COMBREX), and the Critical Assessment of Function Annotation (CAFA), are evaluating and testing the different approaches to predicting the function of proteins of unknown function. These initiatives and global collaborations will increase the capability and reliability of methods to predict biochemical function computationally and will add substantial value to the current volume of structural genomics data by reducing the number of absent or inaccurate functional annotations.

Contents

* Corresponding author at: 360 Huntington Ave., Boston, MA 02115, United States.
E-mail addresses: p.beuning@neu.edu (P.J. Beuning), m.ondrechen@neu.edu (M.J. Ondrechen).

http://dx.doi.org/10.1016/j.csbj.2015.02.003

1. Introduction

The number of protein sequences and structures in databases such as UniProt [1] and the Protein Data Bank (PDB) [2] has grown significantly since the inception of genome sequencing and high-throughput structure determination. As of January 2015, the UniProt/TrEMBL database contains over 89 million protein sequence entries, an increase of more than six-fold since January of 2011; only a very small fraction of these proteins is assigned a reliable function [3]. Additionally, the PDB now includes more than 13,000 structural genomics (SG) protein structures as a result of structural genomics projects, notably the Protein Structure Initiative (PSI). At the turn of the millennium, the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) in the United States launched the PSI with the goal to determine three-dimensional structures of proteins representing every family [4,5]. At that time, the human genome project and the sequencing of the genomes of many other organisms were completed [6,7]. The high throughput techniques developed by the PSI and other SG programs have increased the number of known protein structures. Since the PSI has been primarily concerned with high volume structure determination and prompt public availability of protein structures, most of these protein structures lack reliable accompanying information regarding their biochemical function; in some cases, no functional annotation is given. Thus, most of these proteins are assigned a putative or possible function based on the closest sequence or structure match; however, these assignments are often incorrect [8–10], and these incorrect functional labels can propagate within databases [11,12].

In 2010, the NIGMS launched a new phase of the PSI named PSI:Biology. This phase was implemented to determine the biological roles of the SG proteins under structural study. However, large numbers of functional annotations remain missing or incorrect. Better computational methods and verification through biochemical experimentation are clearly needed. Reliable and accurate computational methods for predicting the function of proteins can add significant value to genomics data and also improve efficiency of experimental verification of function. While there have been a number of review articles on sequence-based and three-dimensional-structure-based methods for function prediction [13–19], this article focuses on newer, local-structure-based computational methods to predict protein function at the molecular level; these methods are in turn based on prediction of the local spatial regions that are biochemically active in the structure. Finally, efforts within the broader scientific community to contribute to the testing and verification of functional predictions are explored.

When the function of a protein is not known, a putative function is sometimes assigned. These assignments are often the result of simple bioinformatics analyses including sequence and three-dimensional structure comparisons using programs such as BLAST [20,21] and Dali [22,23]. SG proteins can be assigned a putative function based on simple transfer of function from the closest sequence or structure match. However, sequence or structural similarities can be misleading. For instance, less than 30% of pairs of proteins with greater than 50% sequence identity have identical E.C. numbers [9]. Even a BLAST E value of 10^{-50} or less does not guarantee that two proteins have the same function [9]. Sequence identities of 60% or greater will transfer function incorrectly in 10% of cases [10]. Furthermore, structural superfamilies, such as the enolase, amidohydrolase, and Clp/crotonase [24] superfamilies, can consist of several, or even dozens, of different biochemical functions [25–28]. The TIM barrel and the Rossmann fold each represent over 50 different types of biochemical function; the TIM barrel has been observed in five out of the six major E.C. categories and the Rossmann fold occurs in all six [29–32]. Thus, the practice of assigning function using simple transfer of function based on sequence or structure similarity has caused misannotations. In one study, the GenBank NR [33], UniProtKB/TrEMBL [1], and Kyoto Encyclopedia of Genes and Genomes (KEGG) [34] databases were shown to have up to 63% misannotation across six superfamilies [8].

For many SG proteins, possible functional assignments obtained from informatics-based approaches can provide too many options with insufficient discrimination of the most likely functions to be able to assign function with confidence or test function experimentally with reasonable efficiency. The development and implementation of new, reliable computational methods is an important aspect of a solution to the challenge of assignment of function to proteins.

2. Functional site prediction methods

Many computational programs have been developed to help predict the active sites and biochemical functions of proteins [16,18,19,35–39], although there remains much yet to be done to improve and to verify predictive capability for biochemical function.

2.1. Sequence-based methods

Sequence-based approaches are the more commonly used method of computational analysis [18]; these methods primarily utilize sequence alignments but sometimes also incorporate 3D structures [40,41]. Evolutionary Trace [42] and INformation-theoretic TREe traversal for Protein functional site IDentification (INTREPID) [43,44], examine a protein in its phylogenetic context and the evolutionary history of each amino acid in a protein sequence to assign a score to each amino acid. Evolutionary Trace analyzes the conservation of residues between proteins of similar function and evaluates amino acid variations that are known to be associated with changes in function. This information then suggests which residues are important for specific functions and which residues can be altered in order to change the function of a protein. This method exploits the similarities and differences between groups of homologous proteins and includes functional resolution, which involves analyzing the different functional clusters that are generated within a given family. Similar to Evolutionary Trace, INTREPID computes scores depending on the degree of conservation within a set of proteins with known functions. This score examines information over an entire family tree instead of just analyzing certain branches, or subfamilies. INTREPID is also able to identify residues important for catalysis that are not necessarily conserved across an entire family [44]. Both methods compute a score for each residue that is a measure of its importance to the function.

2.2. Structure-based methods

Structure-based methods of predicting protein function involve analyzing the structure and shape of a protein. This analysis helps determine where a ligand may bind by transferring the function of another similar protein of known function. Identification of the local site of biochemical activity in a protein can serve as a first step toward the prediction of the function. Geometric-based computational programs like Surfnet [45], CASTp [46], Ligsite [47], PocketFinder [48], and geometric potential [49] are structure-based approaches that examine the different properties of a protein surface or active site pocket to gain insight into the identity and location of binding pockets. Surfnet generates many protein surfaces, such as pockets within a protein, gaps between molecules, and van der Waals interactions, based on PDB coordinate data. The different surface output data are shown as a grid depicting the densities. These grids are created by applying a Gaussian function to the atoms within the protein. The residues that are specified as important are determined from the intensity of the densities. CASTp locates voids within a protein structure using the PDB, Swiss-Prot, and Online Mendelian Inheritance in Man (OMIM) to determine active site residues. Similarly, Ligsite uses a set of ligand–receptor complexes to locate pockets on a protein surface and can analyze a large set of proteins rather quickly [50]. PocketFinder and geometric potential also analyze the topological and geometric features of the protein surface. However, PocketFinder locates ligand binding envelopes instead of scanning the

surface of a protein to find different sized pockets. Geometric potential adds local structural analysis in parallel with global structural analysis to analyze the residues within the pockets.

In addition to geometry-based methodologies, docking methods are another type of structure-based approach. Docking approaches such as Q-SiteFinder [51] and computational solvent mapping [52] identify the position and properties of the catalytic site regions within proteins through the use of small molecule probes. Q-SiteFinder exploits the energy differences between spaces in a protein and van der Waals probes. This helps find the locations in a protein that are energetically favorable for ligand binding. Similarly, solvent mapping uses small organic molecule probes to analyze a protein surface, locates favorable areas where the probes may bind, and then ranks the positions based on their free energies. These methods help locate both catalytic sites and non-catalytic, small molecule binding sites, such as allosteric sites within a given protein structure.

THEoretical Microscopic Anomalous TItration Curve Shapes (or THEMATICS) [53–55], a functional site prediction method, is able to predict accurately the ionizable active site residues within a given protein using only the 3D structure of the query protein. THEMATICS identifies ionizable amino acid residues (Arg, Asp, Cys, Glu, His, Lys, and Tyr, plus the N- and C- termini) that participate in catalysis or ligand recognition. The ionizable side chains of amino acid residues in protein active sites exhibit unusual electrostatic properties, specifically theoretical titration curves as shown in Fig. 1. These curves are obtained by approximate calculation of the electrostatic potential function, followed by a calculation of the average charge of each ionizable residue as a function of pH. These theoretical titration curves of active site residues are perturbed from the normal sigmoidal shape that is characteristic of the Brönsted acid–base chemistry of the free amino acid [53]. In a normal titration curve, the proton occupation is one at low pH and as the pH is increased, the proton occupation suddenly drops sharply around the pK_a, approaching zero at higher pH. Normally this transition, where both the protonated and deprotonated forms exist in appreciable population, occurs in a narrow pH range. However, the residues within the active site tend to be partially protonated over a larger pH range and in this manner the shape of the titration curve is perturbed [53]. This method has been described previously as based on computed pK_a shifts [38,56]; however, this is incorrect. Only metrics that characterize the shape of the titration curves, and not the pK_a shifts, are used in the THEMATICS predictions. The degree of deviation of a catalytic ionizable residue from the typical Henderson–Hasselbalch titration curve can be quantified by the moments of the first derivative of the curve [57]. This method has been tested on the Catalytic Site Atlas (CSA) 100, and THEMATICS-predicted residues have been shown to constitute good predictions of the active site for proteins in the benchmark set [55]; they have also been shown to be generally well conserved [58].

2.3. Combined methods

In order to take advantage of the strengths of each approach to improve the performance of active site predictors, many current methods utilize structure and sequence-based properties in parallel [59–67]. ConCavity [68] is one method that utilizes both sequence and structure information to predict the functionally active residues of a protein. It uses algorithms that analyze not only the surface of a protein for binding pockets, but also uses evolutionary conservation to help locate these pockets. First, ConCavity scores areas on the surface of a protein according to the topology, using methods such as Ligsite or Surfnet (mentioned above). It then combines the conservation scores of residues within these pocket areas. Next, pocket structures are constructed based on the analysis and the structure of the protein. Finally, the potential pockets are mapped on the structure and the residues are analyzed and scored based on their position with respect to the pockets. With this information, ConCavity is able to predict spaces within a protein structure where a ligand is most likely to bind. The creators of ConCavity have shown that combining structure and sequence analyses significantly improves the ability to identify active site pockets and the residues responsible for catalysis [68].

Since THEMATICS can only predict the seven ionizable amino acids, machine learning methods were developed that can extract more information from the computed electrostatic and chemical properties and can predict all 20 amino acid types. The ionizable residues arginine, aspartate, cysteine, glutamate, histidine, lysine, and tyrosine make up about 76% of active site residues within functionally annotated proteins in databases [69]. To predict all 20 amino acid residue types, a new machine learning method was developed that can analyze the non-ionizable residues as well. This led to the development of Partial Order Optimum Likelihood, or POOL. POOL, a machine learning method, is a maximum likelihood, monotonicity-constrained multidimensional isotonic regression method that has the ability to identify both ionizable and non-ionizable active site residues [70]. POOL accepts THEMATICS metrics for the ionizable residues as one of its input features. However, it also calculates environment variables for all residues based on the THEMATICS metrics for the ionizable residues in the neighborhood of each residue. POOL can accept other input features, including scores from INTREPID [43,44] and the structure-only version of ConCavity [68]. Using structure-based geometric features, ConCavity supplies a score for each residue based on its likelihood of binding to a ligand.

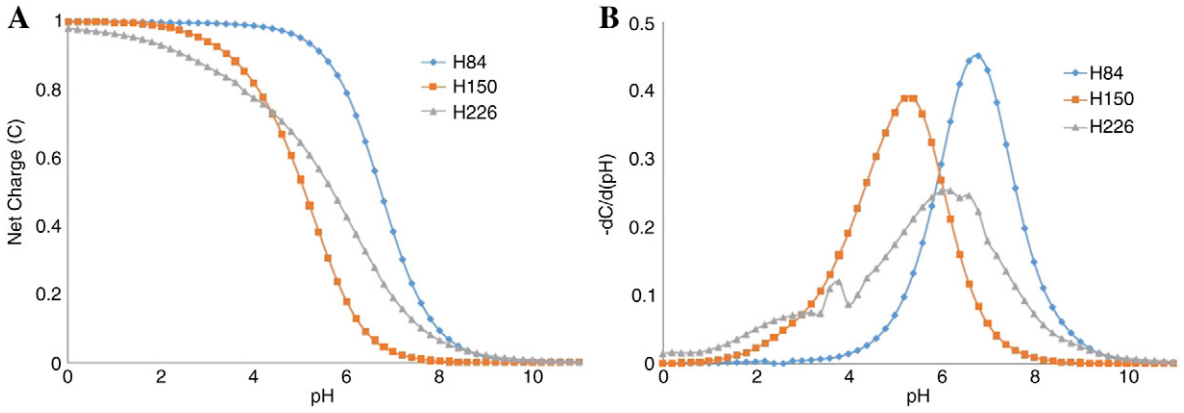

Fig. 1. Three histidine residues from histidinol phosphate phosphatase (HPP) (PDB 2yz5) were analyzed by THEMATICS to produce theoretical titration curves (A), which plot the mean net charge of a given residue of a large ensemble of protein molecules as a function of pH, and first derivative plots (B).The titration curves of two non-catalytic residues, H84 and H150, show sigmoidal curve shapes with a small buffer range, while the catalytic H226 displays a curve with an anomalous shape, shallow slope, and larger buffer range. When analyzing the first derivatives of the titration curves, non-catalytic residues display symmetrical, highly peaked plots. However, active site residues such as H226 shown here display broad, asymmetric derivative plots and may have multiple peaks.

Together, these three input types from THEMATICS, INTREPID, and structure-only ConCavity generate POOL rankings that yield predictions of the residues that are important for catalysis.

For instance, THEMATICS and POOL were used to analyze the Structural Genomics protein *Bifidobacterium adolescentis* YP_910028.1 of unknown function and predicted that it is a metal-dependent phosphoesterase [71]. Sequence and structure comparisons with BLAST and Dali were inconclusive and suggested multiple different functions. The closest structure match was to a DNA polymerase catalytic domain. Initial phylogenetic analysis suggested that this protein could function to repair DNA or function as a DNA polymerase.

The crystal structure of YP_910028.1 contains a PHP domain, but PHP domains are present in multiple functional types, including X-family DNA polymerases [72], DNA polymerase III [73], and a histidinol phosphate phosphatase [74]. The location of the iron and zinc metals can suggest a general location for the active site, but cannot be used to determine a specific function since these trinuclear metal-binding sites are seen in a range of diverse proteins including endonucleases, phosphatases, and phospholipases [75–78]. Other analyses [79,80] were unable to provide a definitive functional annotation.

THEMATICS and POOL analysis of YP_910028.1 predicted sets of residues that closely match those predicted for histidinol phosphate phosphatase (HPP, PDB ID 2yz5) in a local structure alignment, with weaker matches to the other proteins of known function with similar folds, suggesting phosphoesterase activity for the enzyme. DNA polymerase III (PDB: 2hpi) has a similar metal-binding motif, but key cysteine and tyrosine residues are replaced by histidine and threonine residues in YP_910028.1, respectively. When YP_910028.1 is superimposed with both DNA polymerase III and HPP, the predicted active site residues align better with HPP (Fig. 2). This indicates that YP_910028.1 possesses phosphoesterase activity and not DNA polymerase activity. Phosphoesterase activity was detected by observation of the hydrolysis of the phosphate group of *para*-nitrophenyl phosphate (pNPP) to form *p*-nitrophenol and was shown to be dependent on the concentration of YP_910028.1. However, the tests for DNA polymerase activity resulted in no detectable activity regardless of the conditions used [71].

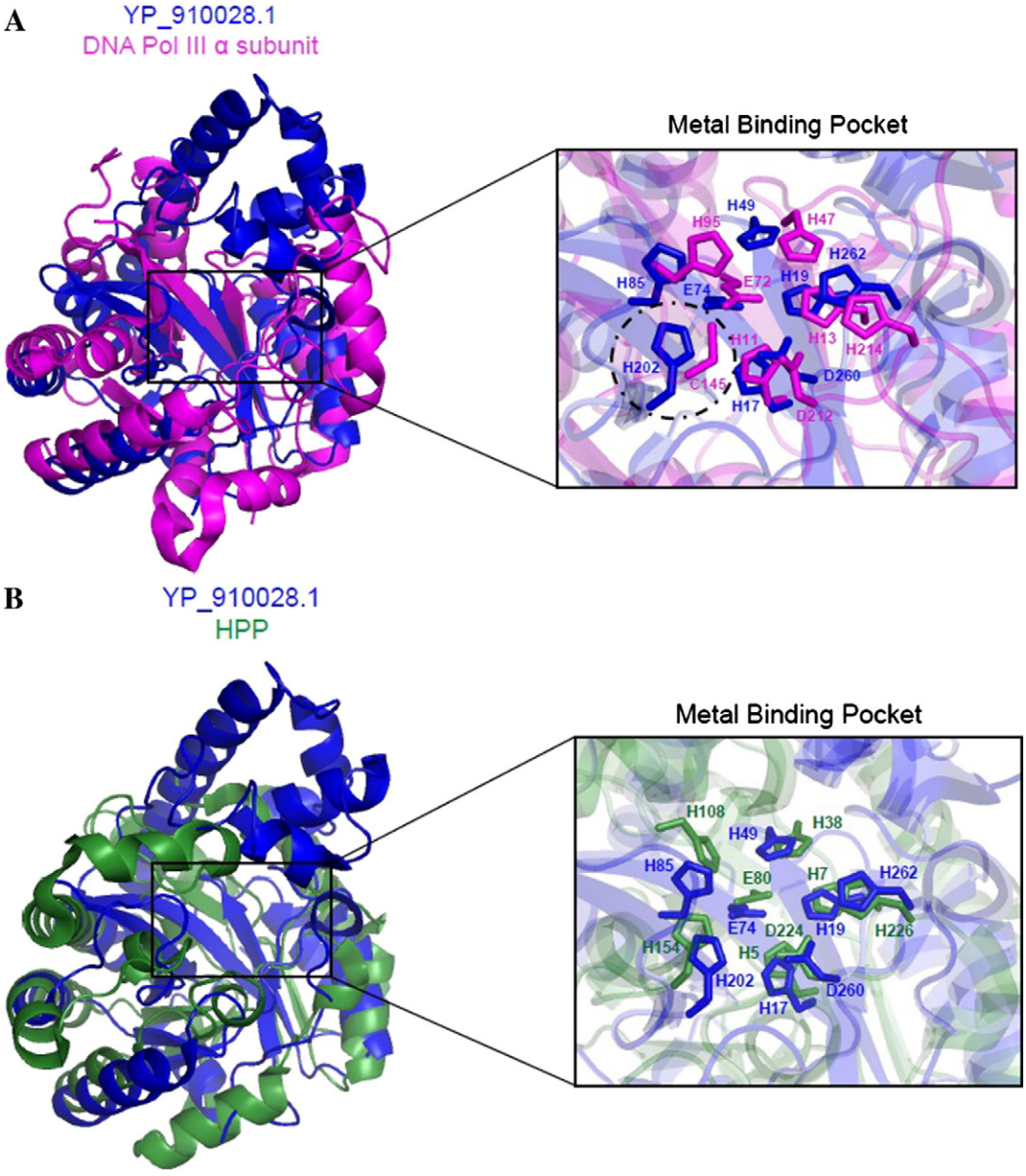

Fig. 2. (A) The metal binding pocket of YP_910028.1, containing a PHP (Polymerase and Histidinol Phosphatase) domain (PDB ID 3e0f, shown in dark blue) aligns well with that of DNA Pol III alpha subunit (PDB ID 2hpi, shown in magenta). However, C145 and Y74 of DNA Pol III are mismatched with a histidine and threonine, respectively in YP_910028.1. (B) On the other hand, the metal binding pocket of YP_910028.1 (PDB 3e0f) aligns perfectly with the pocket of histidinol phosphate phosphatase (HPP) (PDB 2yz5), shown in green.

3. Annotating protein function

3.1. Local active site prediction methods

In comparison to global sequence- and structure-based methods that analyze an entire protein, local active site prediction methods find the biochemically active local region of the structure and then focus on the residues within the pocket and in the immediate surroundings. These methods are useful when analyzing entire families of proteins for which a specific signature is observed within the local active site.

For example, ProBiS [81] is a web server that utilizes an algorithm to detect similarities within protein binding pockets through local structural alignments of multiple proteins. ProBiS provides access to a database of 420 million pairwise local structure alignments and will perform pairwise local alignments for structures that are not in its database.

3.1.1. ProFunc

ProFunc [82] is a metaserver that combines sequence, global structure, and local structure-based methods to obtain a set of function predictions from which one might seek consensus. First, the protein of unknown function is analyzed by numerous sequence searches, shown on the left-hand side in Fig. 3. BLAST [20,21] analysis scans both the PDB and UniProt and uses multiple sequence alignment to determine sequence similarities and detect sequence motifs [83]. Gene neighbors are also examined based on the query protein's predicted location within the genome. The genes located near each other are often functionally related or functionally similar [82]. Next, structure-based analyses are performed on the protein of interest. This involves searching a number of databases for global folds or local structures that are similar to the query protein. Surfnet, mentioned in the above section, is one of these databases. Another database, secondary structure matching (SSM) [84] evaluates the secondary structure elements (SSEs) of the query protein of unknown function and compares them to the SSEs of protein structures within its database. The algorithm retrieves high, strong matches and superimposes the structures with the query protein to give a root mean square deviation (RMSD) so that a common number can be used to compare the results. Finally, ProFunc utilizes other servers to search for 3D templates of proteins with known binding sites. These binding sites may be simple active sites with the residues important for catalysis known [85], or ligand binding sites wherein residues important for catalysis are known and also the natural ligand/substrate is known. In some cases, the databases can also compare DNA-binding sites and motifs known to be associated with binding DNA.

3.1.2. Structurally Aligned Local Sites of Activity (SALSA)

The computational method Structurally Aligned Local Sites of Activity, or SALSA [86] utilizes a combination of functional residue prediction from POOL with local three-dimensional structural alignments. The characteristic spatial patterns of predicted residues at the local active site are used to identify biochemical functions. For example, a superfamily can consist of a number of functional families, each with a biochemical function that is different from the other members of that superfamily. SALSA tables can be constructed using the locally aligned residues at the predicted active sites across the entire superfamily. Proteins with the same function generally possess a particular spatial pattern or signature of predicted functional residues, while proteins of different functions possess different signatures. This consensus signature for each functional family is established using POOL predictions for a set of proteins with known common function; this defines the signature for each of the known functional types within a superfamily. If the superfamily contains SG proteins, the predicted sets of functional residues for the SG proteins can be compared with the consensus signatures for the known functional families. Thus, SALSA defines the different kinds of active sites, and therefore different functional types, within a superfamily. The general method is illustrated in the workflow shown in Fig. 4.

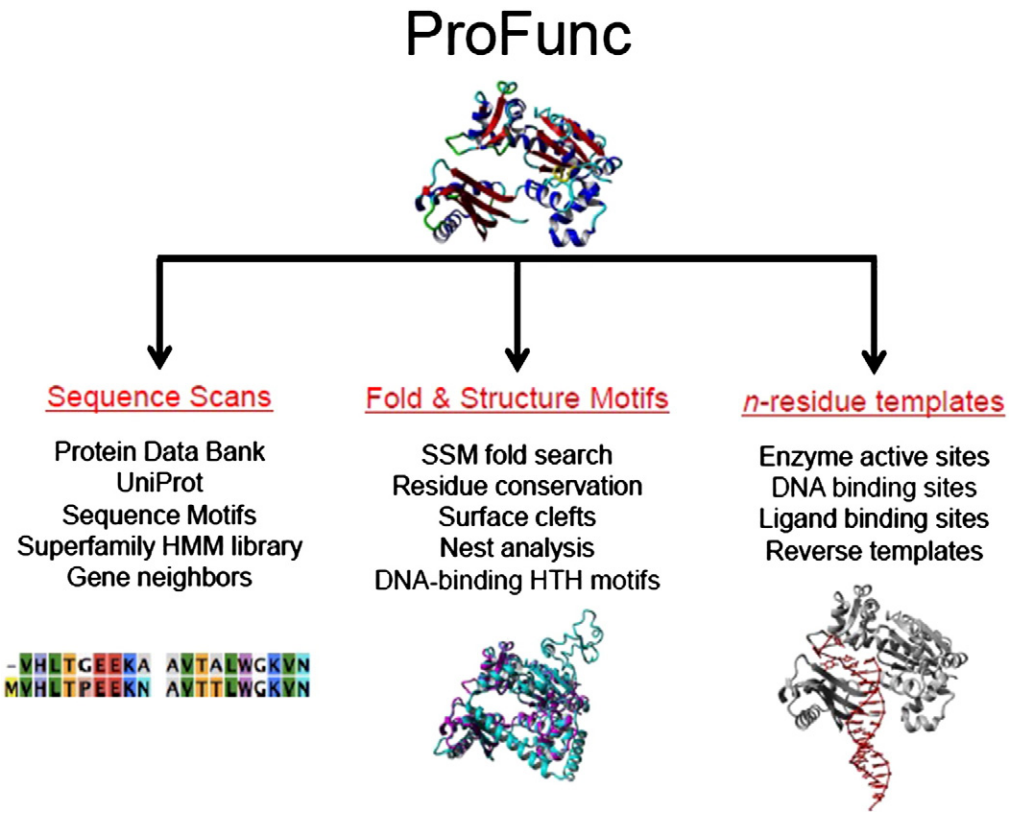

Fig. 3. Schematic diagram outlining the different methods utilized in ProFunc. HMM: Hidden Markov Model; SSM: Secondary Structure Matching; HTH: Helix–Turn–Helix.

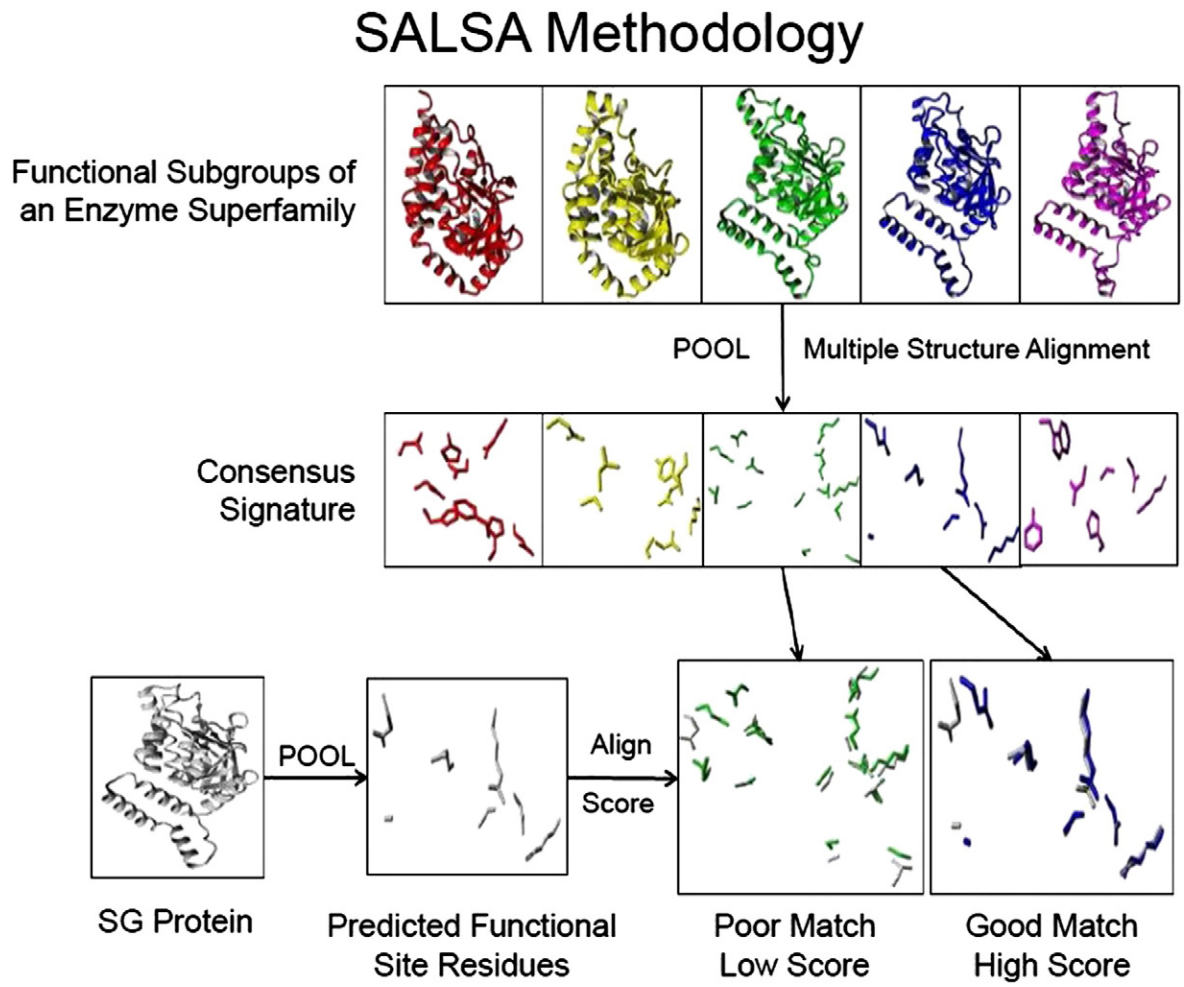

Fig. 4. Schematic diagram outlining the SALSA method of annotating protein function.

3.2. Community initiatives and projects

In an effort to tackle the growing challenges of protein function prediction and the correction of enzyme function misannotations within databases, the community has come together to take on the challenge. These global projects involve collaboration between numerous groups, employing theory, computation, and experiment, and have started to make significant progress toward the confirmation of protein function, thus adding a substantial value to the information on structural genomics proteins currently available.

3.2.1. The Enzyme Function Initiative (EFI)

The Enzyme Function Initiative (EFI) [3], funded by NIGMS, began 10 years after the start of the PSI. This initiative combines bioinformatics with experimental enzymology to help determine the substrate specificity of proteins of unknown function. Each aspect of the EFI can be divided into whether or not the work can be done in a high throughput, moderate throughput, or low throughput manner. Generally, the first steps of the project, computational and bioinformatics analysis, fall under high throughput methods that help focus the experimental work in the final stages of this project, which involve lower throughput methods. The initial bioinformatics analyses, including database searches for sequences and structures of unknown function, preliminary molecular ligand docking, and clustering of pathways, can be executed on a high throughput basis [87]. Experimental enzymology, including preliminary homology modeling, expression and purification of enzymes of interest, and screening enzymes for different activities can be done at a rate of a few enzymes per month and falls under moderate throughput. The limiting factors of this project, however, are the experiments that fall under the low throughput category, including obtaining higher resolution homology models and docking studies, determining structure–function relationships, in vivo studies of functional predictions, and identification of enzymes with functional promiscuity [88,89], each of which can be highly demanding of time and labor. However, the preliminary work helps refine the experimental analysis, which highlights the necessity of reliable computational prediction methods to be used in parallel with experimental validation methods.

The project focuses these methods on five superfamilies with diverse functions that have been selected as test cases for developing the strategy outlined above: (1) amidohydrolase (AH), (2) enolase (EN), (3) glutathione transferase (GST), (4) haloalkanoic acid dehalogenase (HAD), and (5) isoprenoid synthase (IS). These Bridging Projects help determine target enzymes as well as information about the enzymes of known function in each superfamily.

In order to be successful, the EFI strategy must be able to assign a novel function for enzymes that are functionally diverse from enzymes of known function. However, molecular docking of a ligand into an enzyme is not always a reliable way to determine substrate specificities. In particular, substrates can cause conformational changes in vitro that are not observed in silico and the scoring algorithms may not be accurate [3]. At the end of its term, the EFI proposes that it will have a working strategy consisting of a set of databases and programs that the scientific community can utilize in expanding this analysis to every protein superfamily.

This method has been successfully tested on numerous proteins of unknown function. Specifically, the in silico docking method of the EFI described above has been successfully applied to the entire dipeptide epimerase family within the EN superfamily. Within this superfamily, a member of the *cis,cis*-muconate lactonizing enzyme (MLE) family encoded by the *Bacillus cereus* ATCC 14579 genome with previously unknown function was predicted to have *N*-succinyl arginine racemase function based on docking approaches [90]. A virtual library consisting of *N*-succinyl amino acids and dipeptides was virtually docked into a homology model of this enzyme. The homology model was created using a series of template structures from the PDB. The structure of L-Ala-D/L-Glu epimerase from *Bacillus subtilis* (PDB ID 1TKK) was the template that contributed the most to the homology model. This template was also prominent in many subsequent homology models for members of the dipeptide epimerase family and was useful in the docking studies of nearly 700 enzymes.

Another successful docking study, performed by one of the Bridging Projects, aided in assigning function to *Thermotoga maritima* Tm0936, a member of the AH superfamily whose function was previously unknown. Tm0936 was predicted to have a novel function as an *S*-adenosylhomocysteine deaminase [91]. This study involved docking thousands of metabolites into Tm0936 and creating a target list comprising adenine analogues. Five potential substrates were chosen based on availability and rank within the docking study; of these, the enzyme had significant activity with three: adenosine, 5-methylthioadenosine (MTA), and *S*-adenosylhomocysteine (SAH) (Fig. 5). It was concluded that this enzyme is involved in the deamination of metabolites within the MTA/SAH pathway.

3.2.2. Critical Assessment of Function Annotation (CAFA) experiment

Until recently, there was no way to compare the performance of different automated function prediction methods. Over the past few years, Iddo Friedberg and Predrag Radivojac, through collaboration with many computational research groups, have designed an experiment to test multiple automated function prediction tools and programs. This Critical Assessment of Function Annotation (CAFA) [92] experiment is a large-scale community-wide collaboration designed to evaluate the performance of the many diverse methodologies [60,82,93–99] developed by research groups over the years. These methods range from studying protein–protein interactions [100–103] to analyzing sequences [104–108] to examining evolutionary features of proteins [109–113]. The main focus is to evaluate the quality of current sequence-based automated function prediction methods and to identify the computational methods that perform the best in predicting correct or novel functions.

So far, the CAFA experiment has gone through two experimental periods, with the second experiment recently completed. In both instances, the protocols, or "rules," are similar. The classification system used by the CAFA experiments was developed based on the definition of protein function classification by the Gene Ontology (GO) Consortium [114]. The GO project utilizes many different databases [115–142] to help provide a solution to the problem of automated function prediction. The main goal of the GO Consortium is to develop a uniform vocabulary to use when describing the functions of all eukaryotic proteins. The first CAFA project lasted 15 months and consisted of 30 teams of researchers from around the globe, who tested over 50 algorithms designed to annotate protein function. The different methods were tested on a set of over 860 protein sequences spanning 11 species, including *Escherichia coli*, *B. subtilis*, and *Homo sapiens* [92].

From the GO Consortium categories, this project involves information from the "Biological Process" and "Molecular Function" sections. These sections are two of the three structured vocabularies that the GO project has developed to describe gene products. The experiment began with providing a set of over 48,000 proteins of uncertain biochemical function to the teams involved. After the teams worked on annotating these proteins, the assessors performed GO experimental annotations over the course of almost a year. Of the protein sequences analyzed, a set of 866 were chosen based on the accumulation of functional annotations made over the 11 month period. The published results revolve around a maximum *F*-measure, also known as F_{max}, which corresponds to a "harmonic mean between precision and recall" [92]. Two methods, BLAST [20,21] and a Naïve baseline method [92], were used to compare the test methods. In the BLAST method, the GO terms that define any protein sequences for which a function has been experimentally determined are assigned to the sequence being analyzed. In the Naïve method, the GO terms used to describe the target sequences are scored based on how frequently the term comes up in the Swiss-Prot database overall.

This large-scale CAFA experiment and others to follow like CAFA2 are designed to help researchers evaluate their methods in comparison to other methods in existence. They also provide the community with a set of predictions for a number of proteins of unknown or uncertain function. Overall, the results of the first experiment showed accurate performances when predicting the "Molecular Function" of the target proteins. However, the same could not be said for predicting the "Biological Processes" of the target proteins, which shows room for improvement in all methods.

The two top performing methods for predicting both "Molecular Function" and "Biological Process" ontologies were Jones—UCL [143] and Argot2 [144]. The Jones—UCL method uses known protein–protein interactions, gene expression, and sequence similarity to assign protein functions [143]. The Argot2 method analyzes a given protein sequence by BLAST [20,21] and HMMer [145,146] first, followed by a search of GO terms from the UniProtKB-GOA database [138]. The results highlight the improvement in function prediction that can be gained from combining multiple input features.

In the first CAFA global project (CAFA1), an analysis of human mitochondrial polynucleotide phosphorylase 1 (hPNPase) from a family of exoribonucleases was reported. This large protein works in complex with other portions of the mitochondrial degradosome and is characterized by a number of diverse functions for which experimental data exist.

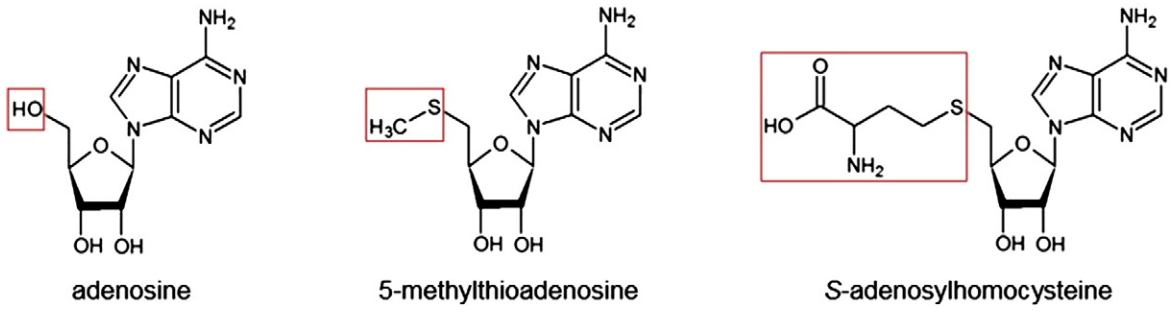

Fig. 5. The metabolites above dock in silico into Tm0936 and are substrates of the enzyme Tm0936. The general structure of these three metabolites is the same with the exception of the moieties shown in the boxes.

These functions include hydrolyzing single-stranded RNA [147], facilitating the import of RNAs into the mitochondrial matrix [148], and responding to oxidative stress [149]. A number of methods under examination in the CAFA project made predictions for hPNPase. In the "Molecular Function" GO terms category, most methods were able to predict correctly two functions for hPNPase: single-stranded RNA hydrolysis and import of small RNAs. Other functions are more uncommon within the family of hPNPase, which may contribute to the lack of methods able to predict these functions [92]. The most well-known biological function of hPNPase is the import of RNA into the mitochondria. Within the "Biological Process" GO terms category, this major function as well as others were not predicted.

3.2.3. COMputational BRidges to EXperiments (COMBREX)

The COMBREX Project's main goal is to understand and annotate the function of microbial proteins [150]. As its name implies, this project brings theorists and experimentalists together in order to increase the rate at which proteins from archaeal and bacterial genomes are functionally annotated [151]. There are three main components to this project: the *COMBREX Community*, the *COMBREX Database*, and the *COMBREX grants*. The grants are used to fund community members working on the efforts described above, while the database serves as a universal place to house the list of functionally annotated proteins. Currently, this database contains more than 3.3 million proteins spanning over 1000 microbial genomes [150]. Of the genes in the database, less than 0.5% have experimental data regarding the function of the gene. However, over 75% of the genes contain a computationally predicted function, but lack experimental validation. In general, the COMBREX project is working toward creating a Gold Standard Database to serve as the basis for training algorithms for future protein annotation methods. During the beginning of the COMBREX-funded projects, experimentalists were assigned 140 proteins on which to perform experiments. Of these 140 proteins, 37 contain 28 unique domains that are similar to human proteins, which potentially can lead to new information about human health and diseases. Also within this group of proteins are eight domains of unknown function defined by Pfam, which allows for some novel predictions of function to be made. Of these 140 proteins, about half have a successfully validated functional prediction [152–156]. In one instance [155], bacterial YbbB is identified in twelve archaeal genomes and its function is determined to be a tRNA 2-selenouridine synthase. In order to confirm this functional classification, first preliminary computational analysis, including BLAST [20,21] searches, was performed on the protein of uncertain function. Next, structure-based alignments and neighboring genes were analyzed using CLUSTAL W [157] and a neighbor-joining method [158]. To validate the results of the computational methods, in vitro activity assays were performed by gene complementation/replacement [159,160] and tRNA selenation [155] experiments. In the end, the computational predictions were successfully validated by the experimental methods, and the function of this protein was determined.

4. Summary and outlook

The process of annotating proteins of unknown and uncertain functions continues to be challenging yet critical for understanding the enormous amount of information generated by genome sequencing and structural genomics projects. Function prediction methods that focus on the local spatial region of biochemical activity show promise for improving predictive capability. Proteins that contain high sequence similarity on a global level do not always have that same sequence similarity at the local active site. Conversely, proteins with low overall sequence similarity can have high similarity in the spatial region of the active site. Too often, the function of a protein that has high global sequence similarity with a protein of unknown function is transferred to the target protein without analyzing the local active site sequence similarities.

In an effort to provide useful information about enzymes of unknown function, many research groups have developed methods to predict protein function. However, the probability of misannotation is higher when only one type of analysis, sequence- or structure-based, is used when making predictions. As methods continue to be optimized and used in parallel with other methods, the information obtained though the genome projects can become more useful and complete. With the help of these breakthrough computational methods listed above and others to come in the future, the challenges of assigning functions to proteins can begin to be resolved. Even with the number of methods available today to predict the function of proteins, it is clear that the field of protein function prediction will continue to grow, especially as the quality and quantity of data continue to increase. While these computational methods are being optimized, biochemical studies can be used to validate the predictions made. Such experimental verification is a major current need in the field. In the future, as computational methods improve and are subjected to experimental verification, biochemical studies can be more focused and less time consuming. Future automation of the computational methods will enable fast, high-throughput functional annotation of these proteins and thus add significant value to the vast, growing store of genomics data.

Acknowledgments

Support of the National Science Foundation under grant number CHE-1305655, a grant from MathWorks, an American Cancer Society Research Scholar Grant, RSG-12-161-01-DMC (PJB), and a PhRMA Foundation Fellowship (CLM) are gratefully acknowledged.

References

[1] The UniProt Consortium. Activities at the universal protein resource (UniProt). Nucleic Acids Res 2014;42:D191–8.
[2] Westbrook J, Feng Z, Chen L, Yang H, Berman HM. The Protein Data Bank and structural genomics. Nucleic Acids Res 2003;31:489–91.
[3] Gerlt JA, et al. The enzyme function initiative. Biochemistry 2011;50:9950–62.
[4] Stevens RC. Design of high-throughput methods of protein production for structural biology. Structure 2000;8:R177–85.
[5] Burley SK, Joachimiak A, Montelione GT, Wilson IA. Contributions to the NIH–NIGMS Protein Structure Initiative from the PSI production centers. Structure 2008;16:5–11.
[6] Venter JC, et al. The sequence of the human genome. Science 2001;291:1304–51.
[7] Lander ES, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
[8] Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 2009;5:e1000605.
[9] Rost B. Enzyme function less conserved than anticipated. J Mol Biol 2002;318:595–608.
[10] Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 2003;333:863–82.
[11] Gilks WR, Audit B, de Angelis D, Tsoka S, Ouzounis CA. Percolation of annotation errors through hierarchically structured protein sequence databases. Math Biosci 2005;193:223–34.
[12] Llewellyn R, Eisenberg DS. Annotating proteins with generalized functional linkages. Proc Natl Acad Sci U S A 2008;105:17700–5.
[13] Gabaldon T. Computational approaches for the prediction of protein function in the mitochondrion. Am J Physiol Cell Physiol 2006;291:C1121–8.
[14] Pandey G, Kumar V, Steinbach M. Computational approaches for protein function prediction: a survey, TR 06-028. Twin Cities: Department of Computer Science and Engineering, University of Minnesota; 2006.
[15] Hu P, Bader G, Wigle DA, Emili A. Computational prediction of cancer-gene function. Nat Rev Cancer 2007;7:23–34.
[16] Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 2007;8:995–1005.
[17] Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol 2007;3:88.
[18] Loewenstein Y, et al. Protein function annotation by homology-based inference. Genome Biol 2009;10:207.
[19] Sleator RD, Walsh P. An overview of in silico protein function prediction. Arch Microbiol 2010;192:151–5.
[20] Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
[21] Cameron M, Williams HE, Cannane A. Improved gapped alignment in BLAST. IEEE/ACM Trans Comput Biol Bioinform 2004;1:116–29.
[22] Holm L, Kaariainen S, Wilton C, Plewczynski D. Using Dali for structural comparison of proteins. Curr Protoc Bioinformatics 2006;S14:5.5.1–5.5.24 [Chapter 5: Unit 5 5].

[23] Holm L, Rosenstrom P. Dali server: conservation mapping in 3D. Nucleic Acids Res 2010;38:W545–9.
[24] Hamed RB, Batchelar ET, Clifton IJ, Schofield CJ. Mechanisms and structures of crotonase superfamily enzymes — how nature controls enolate and oxyanion reactivity. Cell Mol Life Sci 2008;65:2507–27.
[25] Brown SD, Gerlt JA, Seffernick JL, Babbitt PC. A gold standard set of mechanistically diverse enzyme superfamilies. Genome Biol 2006;7:R8.
[26] Hegyi H, Gerstein M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol 1999; 288:147–64.
[27] Furnham N, et al. Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies. PLoS Comput Biol 2012;8:e1002403.
[28] Friedberg I, Godzik A. Functional differentiation of proteins: implications for structural genomics. Structure 2007;15:405–15.
[29] Nagano N, Orengo CA, Thornton JM. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 2002;321:741–65.
[30] Anantharaman V, Aravind L, Koonin EV. Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. Curr Opin Chem Biol 2003;7:12–20.
[31] Kinoshita K, Nakamura H. Protein informatics towards function identification. Curr Opin Struct Biol 2003;13:396–400.
[32] Omelchenko MV, Galperin MY, Wolf YI, Koonin EV. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. Biol Direct 2010;5:31.
[33] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res 2009;37:D26–31.
[34] Kanehisa M, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res 2008;36:D480–4.
[35] Chi X, Hou J. An iterative approach of protein function prediction. BMC Bioinformatics 2011;12:437.
[36] Wilkins AD, Bachman BJ, Erdin S, Lichtarge O. The use of evolutionary patterns in protein annotation. Curr Opin Struct Biol 2012;22:316–25.
[37] Watson JD, Laskowski RA, Thornton JM. Predicting protein function from sequence and structural data. Curr Opin Struct Biol 2005;15:275–84.
[38] Gherardini PF, Helmer-Citterich M. Structure-based function prediction: approaches and applications. Brief Funct Genomic Proteomic 2008;7:291–302.
[39] Skolnick J, Brylinski M. FINDSITE: a combined evolution/structure-based approach to protein function prediction. Brief Bioinform 2009;10:378–91.
[40] Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol 1995;2:171–8.
[41] Glaser F, et al. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 2003;19:163–4.
[42] Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–58.
[43] Sankararaman S, Kolaczkowski B, Sjolander K. INTREPID: a web server for prediction of functionally important residues by evolutionary analysis. Nucleic Acids Res 2009;37:W390–5.
[44] Sankararaman S, Sjolander K. INTREPID—INformation-theoretic TREe traversal for Protein functional site IDentification. Bioinformatics 2008;24:2445–52.
[45] Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 1995;13(323–30):307–8.
[46] Dundas J, et al. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 2006;34:W116–8.
[47] Huang B, Schroeder M. LIGSITE: predicting protein binding sites using the Connolly surface and degree of conservation. BMC Struct Biol 2006;6. http://dx.doi.org/10.1186/1472-6807-6-19.
[48] An J, Totrov M, Abagyan R. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics 2005;4:752–61.
[49] Xie L, Bourne PE. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. BMC Bioinformatics 2007;8(Suppl. 4):S9.
[50] Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 1997;15: 359–63.
[51] Laurie AT, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. Bioinformatics 2005;21:1908–16.
[52] Clodfelter KH, Waxman DJ, Vajda S. Computational solvent mapping reveals the importance of local conformational changes for broad substrate specificity in mammalian cytochromes P450. Biochemistry 2006;45:9393–407.
[53] Ondrechen MJ, Clifton JG, Ringe D. THEMATICS: a simple computational predictor of enzyme function from structure. Proc Natl Acad Sci U S A 2001;98:12473–8.
[54] Ko J, Murga LF, Wei Y, Ondrechen MJ. Prediction of active sites for protein structures from computed chemical properties. Bioinformatics 2005;21(Suppl. 1): i258–65.
[55] Wei Y, Ko J, Murga LF, Ondrechen MJ. Selective prediction of interaction sites in protein structures with THEMATICS. BMC Bioinformatics 2007;8:119.
[56] Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. Proteins 2005;61:704–21.
[57] Ko J, et al. Statistical criteria for the identification of protein active sites using theoretical microscopic titration curves. Proteins 2005;59:183–95.
[58] Hildebrand DGC, Yang H, Ondrechen MJ, Williams RJ. High conservation of amino acids with anomalous protonation behavior. Curr Bioinforma 2010;5:134–40.
[59] Gutteridge A, Bartlett GJ, Thornton JM. Using a neural network and spatial clustering to predict the location of active sites in enzymes. J Mol Biol 2003;330:719–34.
[60] Pazos F, Sternberg MJ. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci U S A 2004;101:14754–9.
[61] Cheng G, Qian B, Samudrala R, Baker D. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Res 2005;33:5861–7.
[62] Petrova NV, Wu CH. Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. BMC Bioinformatics 2006;7:312.
[63] Innis CA. siteFiNDER|3D: a web-based tool for predicting the location of functional sites in proteins. Nucleic Acids Res 2007;35:W489–94.
[64] Youn E, Peters B, Radivojac P, Mooney SD. Evaluation of features for catalytic residue prediction in novel folds. Protein Sci 2007;16:216–26.
[65] Alterovitz R, et al. ResBoost: characterizing and predicting catalytic residues in enzymes. BMC Bioinformatics 2009;10:197.
[66] Bray T, et al. SitesIdentify: a protein functional site prediction tool. BMC Bioinformatics 2009;10:379.
[67] Sankararaman S, Sha F, Kirsch JF, Jordan MI, Sjolander K. Active site prediction using evolutionary and structural information. Bioinformatics 2010;26:617–24.
[68] Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 2009;5:e1000585.
[69] Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol 2002;324:105–21.
[70] Tong W, Wei Y, Murga LF, Ondrechen MJ, Williams RJ. Partial order optimum likelihood (POOL): maximum likelihood prediction of protein active site residues using 3D structure and sequence properties. PLoS Comput Biol 2009;5:e1000266.
[71] Han GW, et al. Crystal structure of a metal-dependent phosphoesterase (YP_910028.1) from *Bifidobacterium adolescentis*: computational prediction and experimental validation of phosphoesterase activity. Proteins 2011;79:2146–60.
[72] Nakane S, Nakagawa N, Kuramitsu S, Masui R. The role of the PHP domain associated with DNA polymerase X from *Thermus thermophilus* HB8 in base excision repair. DNA Repair (Amst) 2012;11:906–14.
[73] Barros T, et al. A structural role for the PHP domain in *E. coli* DNA polymerase III. BMC Struct Biol 2013;13:8.
[74] le Coq D, Fillinger S, Aymerich S. Histidinol phosphate phosphatase, catalyzing the penultimate step of the histidine biosynthesis pathway, is encoded by *ytvP* (*hisJ*) in *Bacillus subtilis*. J Bacteriol 1999;181:3277–80.
[75] Hosfield DJ, Guan Y, Haas BJ, Cunningham RP, Tainer JA. Structure of the DNA repair enzyme endonuclease IV and its DNA complex: double-nucleotide flipping at abasic sites and three-metal-ion catalysis. Cell 1999;98:397–408.
[76] Kim EE, Wyckoff HW. Reaction mechanism of alkaline phosphatase based on crystal structures. Two-metal ion catalysis. J Mol Biol 1991;218:449–64.
[77] Omi R, et al. Crystal structure of monofunctional histidinol phosphate phosphatase from *Thermus thermophilus* HB8. Biochemistry 2007;46:12618–27.
[78] Hough E, et al. High-resolution (1.5 A) crystal structure of phospholipase C from *Bacillus cereus*. Nature 1989;338:357–60.
[79] Schreier B, Hocker B. Engineering the enolase magnesium II binding site: implications for its evolution. Biochemistry 2010;49:7582–9.
[80] Parasuram R, Lee JS, Yin P, Somarowthu S, Ondrechen MJ. Functional classification of protein 3D structures from predicted local interaction sites. J Bioinform Comput Biol 2010;8(Suppl. 1):1–15.
[81] Konc J, Janezic D. ProBiS-2012: web server and web services for detection of structurally similar binding sites in proteins. Nucleic Acids Res 2012;40:W214–21.
[82] Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 2005;33:W89–93.
[83] Valdar WS, Thornton JM. Conservation helps to identify biologically relevant crystal contacts. J Mol Biol 2001;313:399–416.
[84] Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 2004;60:2256–68.
[85] Furnham N, et al. The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Res 2014;42:D485–9.
[86] Wang Z, et al. Protein function annotation with Structurally Aligned Local Sites of Activity (SALSAs). BMC Bioinformatics 2013;14(Suppl. 3):S13.
[87] Pieper U, et al. ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 2011;39:D465–74.
[88] Cline MS, et al. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2007;2:2366–82.
[89] Akiva E, et al. The structure–function linkage database. Nucleic Acids Res 2014;42: D521–30.
[90] Song L, et al. Prediction and assignment of function for a divergent *N*-succinyl amino acid racemase. Nat Chem Biol 2007;3:486–91.
[91] Hermann JC, et al. Structure-based activity prediction for an enzyme of unknown function. Nature 2007;448:775–9.
[92] Radivojac P, et al. A large-scale evaluation of computational protein function prediction. Nat Methods 2013;10:221–7.
[93] Costello JC, et al. Gene networks in *Drosophila melanogaster*: integrating experimental data to predict gene function. Genome Biol 2009;10:R97.
[94] Huttenhower C, Hibbs M, Myers C, Troyanskaya OG. A scalable method for integration and functional analysis of multiple microarray datasets. Bioinformatics 2006; 22:2890–7.
[95] Kourmpetis YA, van Dijk AD, Bink MC, van Ham RC, ter Braak CJ. Bayesian Markov Random Field analysis for protein function prediction based on network data. PLoS One 2010;5:e9293.
[96] Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. Science 2004;306:1555–8.

[97] Pal D, Eisenberg D. Inference of protein function from protein structure. Structure 2005;13:121–30.
[98] Sokolov A, Ben-Hur A. Hierarchical classification of gene ontology terms using the GOstruct method. J Bioinform Comput Biol 2010;8:357–76.
[99] Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc Natl Acad Sci U S A 2003;100:8348–53.
[100] Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein–protein interaction data. J Comput Biol 2003;10:947–60.
[101] Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 2003;19(Suppl. 1):i197–204.
[102] Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph—theoretic analysis of interaction maps. Bioinformatics 2005;21(Suppl. 1):i302–10.
[103] Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein–protein interaction networks. Nat Biotechnol 2003;21:697–700.
[104] Clark WT, Radivojac P. Analysis of protein function and its prediction from amino acid sequence. Proteins 2011;79:2086–96.
[105] Hawkins T, Luban S, Kihara D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 2006; 15:1550–6.
[106] Jensen LJ, et al. Prediction of human protein function from post-translational modifications and localization features. J Mol Biol 2002;319:1257–65.
[107] Martin DM, Berriman M, Barton GJ. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 2004; 5:178.
[108] Wass MN, Sternberg MJ. ConFunc—functional annotation in the twilight zone. Bioinformatics 2008;24:798–806.
[109] Enault F, Suhre K, Claverie JM. Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC Bioinformatics 2005;6:247.
[110] Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol 2005;1:e45.
[111] Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief Bioinform 2011;12:449–62.
[112] Marcotte EM, et al. Detecting protein function and protein–protein interactions from genome sequences. Science 1999;285:751–3.
[113] Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 1999;96:4285–8.
[114] Ashburner M, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.
[115] Fey P, Dodson RJ, Basu S, Chisholm RL. One stop shop for everything Dictyostelium: dictyBase and the Dicty Stock Center in 2012. Methods Mol Biol 2013;983:59–92.
[116] Hu JC, et al. PortEco: a resource for exploring bacterial biology through high-throughput data and analysis tools. Nucleic Acids Res 2014;42:12330.
[117] Flicek P, et al. Ensembl 2014. Nucleic Acids Res 2014;42:D749–55.
[118] Monaco MK, et al. Gramene 2013: comparative plant genomics resources. Nucleic Acids Res 2014;42:D1193–9.
[119] Kersey PJ, et al. Ensembl Genomes 2013: scaling up access to genome-wide data. Nucleic Acids Res 2014;42:D546–52.
[120] St Pierre SE, Ponting L, Stefancsik R, McQuilton P. FlyBase 102—advanced approaches to interrogating FlyBase. Nucleic Acids Res 2014;42:D780–8.
[121] Chibucos MC, et al. Standardized description of scientific evidence using the Evidence Ontology (ECO). Database (Oxford) 2014:bau075.
[122] Hunter S, et al. InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 2012;40:D306–12.
[123] Orchard S, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 2014;42:D358–63.
[124] Cerqueira GC, et al. The *Aspergillus* Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. Nucleic Acids Res 2014;42:D705–10.
[125] Winsor GL, et al. Pseudomonas Genome Database: improved comparative analysis and population genomics capability for *Pseudomonas* genomes. Nucleic Acids Res 2011;39:D596–600.
[126] Torto-Alalibo T, et al. Genetic resources for advanced biofuel production described with the Gene Ontology. Front Microbiol 2014;5:528.
[127] van Dam TJ, Wheway G, Slaats GG, Huynen MA, Giles RH. The SYSCILIA gold standard (SCGSv1) of known ciliary components and its applications within a systems biology consortium. Cilia 2013;2:7.
[128] Logan-Klumpler FJ, et al. GeneDB—an annotation database for pathogens. Nucleic Acids Res 2012;40:D98–D108.
[129] Begley DA, et al. The Mouse Tumor Biology Database (MTB): a central electronic resource for locating and integrating mouse tumor pathology data. Vet Pathol 2012;49:218–23.
[130] Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic Acids Res 2014;42:D810–7.
[131] Smith CM, et al. The mouse Gene Expression Database (GXD): 2014 update. Nucleic Acids Res 2014;42:D818–24.
[132] Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 2013;41:D377–86.
[133] Wood V, et al. PomBase: a comprehensive online resource for fission yeast. Nucleic Acids Res 2012;40:D695–9.
[134] Laulederkind SJ, et al. The Rat Genome Database 2013—data, tools and users. Brief Bioinform 2013;14:520–6.
[135] Croft D, et al. The Reactome pathway knowledgebase. Nucleic Acids Res 2014;42: D472–7.
[136] Cherry JM, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res 2012;40:D700–5.
[137] Lamesch P, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 2012;40:D1202–10.
[138] Barrell D, et al. The GOA database in 2009—an integrated Gene Ontology Annotation resource. Nucleic Acids Res 2009;37:D396–403.
[139] Yook K, et al. WormBase 2012: more genomes, more data, new website. Nucleic Acids Res 2012;40:D735–41.
[140] Bradford Y, et al. ZFIN: enhancements and updates to the Zebrafish Model Organism Database. Nucleic Acids Res 2011;39:D822–9.
[141] McCarthy FM, et al. AgBase: supporting functional modeling in agricultural organisms. Nucleic Acids Res 2011;39:D497–506.
[142] Inglis DO, et al. The *Candida* genome database incorporates multiple *Candida* species: multispecies search and analysis tools with curated gene and protein information for *Candida albicans* and *Candida glabrata*. Nucleic Acids Res 2012; 40:D667–74.
[143] Cozzetto D, Buchan DW, Bryson K, Jones DT. Protein function prediction by massive integration of evolutionary analyses and multiple data sources. BMC Bioinformatics 2013;14(Suppl. 3):S1.
[144] Falda M, et al. Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. Bmc Bioinformatics 2012;13:S14.
[145] Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol 2011;7:e1002195.
[146] Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 2011;39:W29–37.
[147] Arraiano CM, et al. The critical role of RNA processing and degradation in the control of gene expression. FEMS Microbiol Rev 2010;34:883–923.
[148] Wang G, et al. PNPASE regulates RNA import into mitochondria. Cell 2010;142: 456–67.
[149] Wu J, Li Z. Human polynucleotide phosphorylase reduces oxidative RNA damage and protects HeLa cell against oxidative stress. Biochem Biophys Res Commun 2008;372:288–92.
[150] Anton BP, et al. The COMBREX project: design, methodology, and initial results. PLoS Biol 2013;11:e1001638.
[151] Roberts RJ, et al. COMBREX: a project to accelerate the functional annotation of prokaryotic genomes. Nucleic Acids Res 2011;39:D11–4.
[152] Chatterjee K, et al. The archaeal COG1901/DUF358 SPOUT-methyltransferase members, together with pseudouridine synthase Pus10, catalyze the formation of 1-methylpseudouridine at position 54 of tRNA. RNA 2012;18:421–33.
[153] Clark TA, et al. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. Nucleic Acids Res 2012;40:e29.
[154] Phillips G, et al. Diversity of archaeosine synthesis in crenarchaeota. ACS Chem Biol 2012;7:300–5.
[155] Su D, Ojo TT, Soll D, Hohn MJ. Selenomodification of tRNA in archaea requires a bipartite rhodanese enzyme. FEBS Lett 2012;586:717–21.
[156] Xu SY, et al. Characterization of type II and III restriction–modification systems from *Bacillus cereus* strains ATCC 10987 and ATCC 14579. J Bacteriol 2012;194:49–60.
[157] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–80.
[158] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–25.
[159] Hohn MJ, Palioura S, Su D, Yuan J, Soll D. Genetic analysis of selenocysteine biosynthesis in the archaeon *Methanococcus maripaludis*. Mol Microbiol 2011;81:249–58.
[160] Lin W, Whitman WB. The importance of porE and porF in the anabolic pyruvate oxidoreductase of *Methanococcus maripaludis*. Arch Microbiol 2004;181:68–73.