J Struct Funct Genomics (2012) 13:111–123
DOI 10.1007/s10969-012-9126-6

Structure- and sequence-based function prediction for non-homologous proteins

Lee Sael · Meghana Chitale · Daisuke Kihara

Received: 12 September 2011 / Accepted: 10 January 2012 / Published online: 22 January 2012

Abstract The structural genomics projects have been accumulating an increasing number of protein structures, many of which remain functionally unknown. In parallel effort to experimental methods, computational methods are expected to make a significant contribution for functional elucidation of such proteins. However, conventional computational methods that transfer functions from homologous proteins do not help much for these uncharacterized protein structures because they do not have apparent structural or sequence similarity with the known proteins. Here, we briefly review two avenues of computational function prediction methods, i.e. structure-based methods and sequence-based methods. The focus is on our recent developments of local structure-based and sequence-based methods, which can effectively extract function information from distantly related proteins. Two structure-based methods, Pocket-Surfer and Patch-Surfer, identify similar known ligand binding sites for pocket regions in a query protein without using global protein fold similarity information. Two sequence-based methods, protein function prediction and extended similarity group, make use of weakly similar sequences that are conventionally discarded in homology based function annotation. Combined together with experimental methods we hope that computational methods will make leading contribution in functional elucidation of the protein structures.

Keywords Computational protein function prediction · Structure-based function prediction · Sequence-based function prediction · Ligand binding pocket comparison · Protein surface shape comparison · Pocket-Surfer · Patch-Surfer · Weakly similar sequences

Abbreviations

PDB	Protein Data Bank
3DZD	3 Dimensional Zernike descriptor
ATP	Adenosine triphosphate
HEM	Heme
NAD	Nicotinamide adenine dinucleotide
FAD	Flavin adenine dinucleotide
BTN	Biotin
F6P	Fructose 6-phosphate
GUN	Guanine
PLM	Palmitic acid
RTL	Retinol
AUC	Area under the curve
ROC	Receiver operator characteristic
EF	Enrichment factor
GO	Gene ontology
PFP	Protein function prediction
ESG	Extended similarity group
AFP-SIG	Automatic Function Prediction Special Interest Group
ISMB	Intelligent System in Molecular Biology
CASP	Critical Assessment of Techniques for Protein Structure Prediction

Lee Sael and Meghana Chitale contributed equally to this article.

L. Sael · M. Chitale · D. Kihara (✉)
Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA
e-mail: dkihara@purdue.edu

L. Sael · D. Kihara
Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA

L. Sael · D. Kihara
Markey Center for Structural Biology, Purdue University, West Lafayette, IN 47907, USA

Introduction

The structural genomics projects worldwide have determined an increasing number of protein tertiary structures over a decade [1–4]. As of writing of this article (September 2011), there are over 9,000 structures in the Protein Data Bank (PDB) [5, 6] that were deposited from the structural genomics projects. Among several objectives of these large-scale efforts, one of the major expectations is that the determined structures provide clues for elucidating evolution and function of the proteins [7, 8]. In fact, function of some targeted proteins in the projects have been elucidated by global structure similarity to known characterized proteins [9–12] or by a combination of structural comparison and other sources, such as identified bound cofactors to the structure [13] or biochemical experimental evidences [14].

Protein function should be ultimately investigated experimentally. There are indeed efforts towards systematic functional screening [15, 16] and also efforts to combine computational and experimental function assignments in a genome-scale [7, 17]. In parallel to such experimental method developments, computational function prediction methods are expected to play an important role in post-structural genomics functional elucidation given the fact that computational methods can quickly screen existing data, which is the basis of transferring function from known proteins. Computational methods alone can often sufficient evidences for inferring function of proteins [9–12]. Also Experiments can be greatly benefited if possible functions of uncharacterized proteins are suggested by computational methods [17].

However, conventional computational methods which transfer function from obvious homologous proteins, such as BLAST [18, 19] or FASTA [20] or domain searches [21], leave many protein structures with unknown function, as evidenced in many structures with unknown function deposited to PDB from the structural genomics projects. Methods that predict function from globally similar proteins work well when highly similar known proteins exist in the database. Protein structures solved by the structural genomics efforts, which are left behind in functional annotation, do not have apparent global similarity to any of known proteins. Therefore, to increase the annotation coverage, it is crucial to develop methods which can use local structure similarity or distantly related proteins.

In this article, we briefly review two avenues of computational function prediction methods, i.e. structure-based and sequence-based methods. The focus is on our recently developed function prediction methods, local structure-based methods and sequence-based methods, which can effectively extract function information from distantly related proteins that are discarded by conventional methods. For more general review of computational function prediction, readers are referred to recent comprehensive reviews [22–24].

Approaches for structure-based function prediction

Computational structure-based function prediction methods transfer function of known proteins to a query protein if the known proteins have global or local structural similarity to the query. Structure-based function prediction methods can be divided into global and local approaches. In principle, global methods aim to find distantly related proteins of the same fold to a query protein that are not detectable by considering the sequence similarity [25–27]. Any methods developed for protein structure comparison can be used for this task, e.g. the Combinatorial Extension method [28], Dali [29], SSAP [30], VAST [31], and COSEC [32]. More recently, moment-based methods, the spherical harmonics and the 3D Zernike descriptors (3DZD), have been applied to describe protein surfaces [33–38]. Moment-based methods describe a protein surface as a series expansion of a 3D mathematical function that represents the protein surface shape. Thus, a structure is compactly represented as a vector of coefficients of the series function, which allows a fast real-time structure database search. We showed that some functional class of proteins, e.g. DNA binding proteins, can be detected by surface comparison using the 3DZD, because the descriptors capture the surface shape similarity that is required for the function of the proteins (e.g. saddle like shape of DNA binding regions in DNA binding proteins) [35]. Although global structure similarity indicates functional similarity of proteins in most of the cases, one needs to keep in mind that there are notable exceptions of "superfolds" [39], which are commonly occurring protein folds that are adopted by various protein families. Thus, conservation of functional residues needs to be confirmed before transferring function to a query protein from a known protein of the same global fold.

In contrast to the global structure comparison methods, local structure-based approaches compare local regions of a query protein to a database of known functional sites, e.g. active sites of enzymes. The Catalytic Site Atlas [40], AFT [41], ASSAM [42], SPASM [43], SURFACE [44], FLORA [45], CavBase [46], and SitesBase [47] search local sites with a set of residues/atoms in a query protein that match with known functional sites. Another method, eF-seek [48], represents a protein surface as a graph with nodes characterized with local geometry and electrostatic potentials, and local regions in a query protein that are similar to known functional sites are sought by a sub-graph matching algorithm. Thornton and her colleagues explored the use of the spherical harmonics in representing and comparing protein pockets [49]. Binkowski et al. developed a method

that characterize a pocket with conserved residues at the pocket [50, 51] and also with its pairwise atom distances [52]. The ProFunc server performs various structure- and sequence-based function predictions for a query protein structure ranging from sequence and structural motif searches, active site identification, and global fold comparison [53]. Proknow is another method that integrates multiple evidences for function prediction, such as sequence and structure similarity and protein–protein interactions [54]. The advantage of the local structure-based methods over the global structure-based methods is that the former could identify functional sites in a query protein even when the query does not have evolutionary closely related proteins in a database. To meet the urgent task of annotating protein structures solved by the structural genomics projects that do not have apparent homology to known proteins, we have recently developed local structure-based methods, Pocket-Surfer [55] and Patch-Surfer [56, 57]. These two methods will be overviewed in the next section.

Pocket-Surfer and Patch-Surfer

Pocket-Surfer [55] and Patch-Surfer [56, 57] predict ligand molecules that bind to a query protein by comparing the geometrical shape and the physicochemical properties of a pocket region of the query protein to a database of known binding pockets. Both methods allow a quick real-time scan of the database of binding pockets since they use a rotational invariant pocket representation, which does not require time consuming pre-alignment of pockets upon comparison. Technically, the rotational invariant representation is achieved using the 3DZD, a rotation invariant series expansion of a 3D function [58, 59]. The difference between Pocket-Surfer and Patch-Surfer is that the former represents a pocket as a single object while the latter represents a pocket with a set of surface patches, each of which captures features of local regions of the pocket surface. Figure 1 illustrates the query process of Pocket-Surfer (Fig. 1a) and Patch-Surfer (Fig. 1b).

The first step of the binding ligand prediction for a query protein by Pocket-Surfer and Patch-Surfer is to generate the surface of the protein from its PDB file and to extract binding pockets from the surface. The surface of the proteins is determined by the boundary of solvent accessible and solvent excluded regions generated by the Adaptive Poisson-Boltzmann Solver (APBS) program [60]. If the center location of binding pockets in the query protein is not known, it can be predicted with an external pocket finding program, such as VisGrid [61] or LIGSITE [62]. The extent of a pocket surface is computed by casting rays from the predetermined pocket centers and selecting the surface positions that are encountered first by the rays as the pocket surface. Once a pocket is determined, Pocket-Surfer encodes the geometrical shape and the surface electrostatic potential of the whole pocket with 3DZDs. On the other hand, Patch-Surfer segments the pocket surface into patches. The segmentation is done by spreading seed points on the pocket surface and extracting local surface regions that are included within a sphere of 5 Å radius centered at each seed point. For example, adenosine

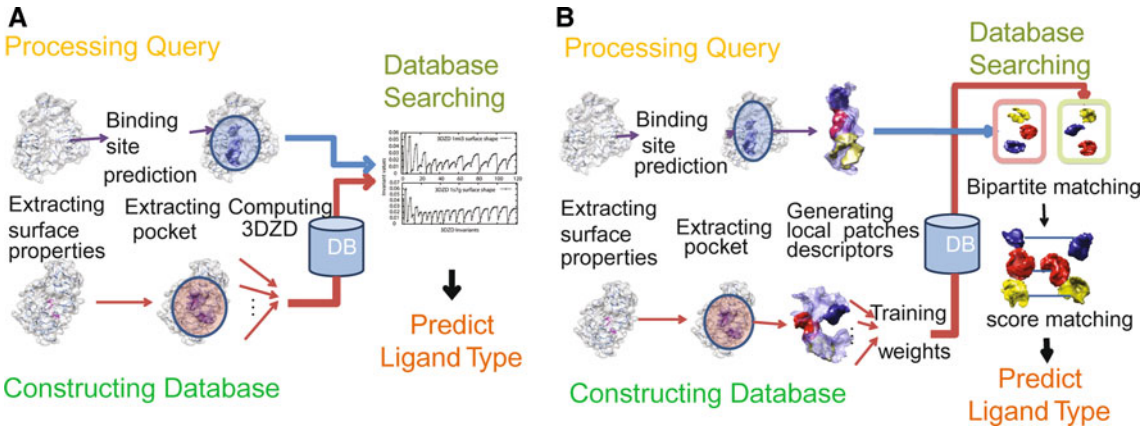

Fig. 1 The flow chart of pocket search with Pocket-Surfer and Patch-Surfer. **a** Pocket-Surfer. Pockets are encoded by the 3DZD. Examples of the 3DZDs are shown in graphs on the *right hand side*. The *x* axis of the plot indicates the position of terms in the series expansion and the *y* axis is the value of the coefficients of the terms. The similarity of two 3DZDs is computed as the Euclidean distance of the two vectors of the 3DZDs. **b** Patch-Surfer. Since a pocket is represented by a set of patches, first, similar patches from the two pockets are matched using a bipartite matching algorithm. The similarity of two pockets reflects the average similarity of the matched patches, the relative position (distance) of patches within each pocket, and the size of the pockets. The weighting factors, which are trained on the known pockets in the database, are used for normalizing scores of different properties at different patches by considering the distribution of the scores

triphosphate (ATP) binding pocket is represented with, on average, 29.5 overlapping patches. Then, the shape, the electrostatic potential, and the hydrophobicity of each patch are mapped on a 3D grid, each of which is considered as a 3D function and encoded with the 3DZD. Since the 3DZD is a vector of coefficients of each term in the series function, the similarity of two 3DZDs can be efficiently quantified by computing the Euclidean distance of the two vectors. Another advantage of the 3DZD is that the series expansion does not change with rotations of the target 3D object (rotationally invariant). Thus, time-consuming pre-alignment of pockets is not needed for comparison. For mathematical details of the 3DZD, refer to the original papers [35, 56, 58, 59].

Once the query pocket is encoded with the 3DZDs, it is compared to the known pockets stored in the database. Pocket-Surfer simply computes the Euclidean distance between the 3DZD of the query pocket and the 3DZD of the database pockets. The comparison of a pair of pockets is more complex for Patch-Surfer since a pocket is represented by a set of patches. Patch-Surfer first identifies pairs of similar patches from the two pockets by employing a modified bipartite matching algorithm [63]. Then, the similarity of the two pockets is quantified by a linear combination of three terms, the average similarity of matched pairs of patches, the relative distance of patches within each pocket, and the size of the pockets. Next, the existing pockets in the database are sorted by the distance to the query pocket. Finally, the top k most similar pockets are considered to make the final prediction of the binding ligand for the query pocket using a type of the k-nearest neighbour algorithm (essentially the algorithm takes weighted consensus within the k highest ranking ligands). Please refer to the original papers for the details of the algorithm [56, 57].

Performance of Pocket-Surfer on benchmark datasets

We benchmarked the accuracy of pocket retrieval of Pocket-Surfer on two datasets of ligand binding pockets selected from PDB. The first dataset contains 100 proteins that bind either one of nine different ligand molecules including ATP, nicotinamide adenine dinucleotide (NAD), flavin adenine dinucleotide (FAD), and glucose. The second dataset contains 175 proteins that bind one of twelve different ligand molecules. There are no overlap between the ligand types and proteins in the two datasets. Pocket-Surfer identified correct ligand types within top 3 ranks 75.6% of the cases for the first dataset and 61.5% for the second dataset. These results were superior to other similar moment-based methods compared in the study [55]. In addition, comparison with existing binding ligand prediction servers showed that Pocket-Surfer achieved the highest value for the area under the Receiver Operator Characteristic curve (AUC-ROC) [33]. Please refer to the original papers [33, 55] for more detailed results of the benchmark studies.

Patch-Surfer results on the representative binding pocket database

Binding pockets of the same ligand type do not always have similar global shape and physicochemical properties at the corresponding location in the pockets [64]. This divergence of properties of pockets can occur due to several reasons: For example, some ligand molecules can take different conformations upon binding. Also occasionally water molecules or additional ligand molecules bind at the same pocket, which results in the change of overall pocket shape, size, and properties.

The intention of the patch-representation by Patch-Surfer is to identify local surface regions that are consistent in shape and/or physicochemical properties in pockets of the same ligand type that do not have globally similar shape and properties. Overall the performance of Patch-Surfer is better than Pocket-Surfer in a benchmark study we conducted on the dataset of 100 proteins that bind to one of nine different ligand molecules [55, 56]. Pocket-Surfer made correct binding ligand prediction for 36.1 and 82.7% of the cases within top-1 and top-3 predictions (i.e. correct ligand is predicted within top-1/top-3 highest scoring ligands ranked by Pocket-Surfer), whereas Patch-Surfer's results were 45.0 and 86.0% for the top-1 and top-3 predictions, respectively. The area under the curve (AUC) value of the receiver operator characteristic (ROC), a metric to evaluate the overall database retrieval performance [65], was 0.81 for Pocket-Surfer while 0.82 is achieved by Patch-Surfer [56].

We have recently developed a larger database of representative ligand binding pockets selected from PDB for practical use of Patch-Surfer [66]. The representative pockets were selected from the Protein-Small-Molecule DataBase (PSMDB) [67]. Among several non-redundant datasets of structures of protein–ligand complexes provided in PSMDB, we chose the list available at http://compbio.cs.toronto.edu/psmdb/downloads/CPLX_25_0.85_7HA.list , where proteins were pruned with 25% sequence identity and redundant ligands that have a Tanimoto coefficient of 0.85 or higher to other ligand molecules were filtered out. Small ligands with less than 7 heavy atoms were not included in this list. From this list, we further removed ligands that are too distant from the protein (more distant than 3.5 Å to any heavy atom in the protein) and also covalently bound ligands (ligands that are closer than 1.4 Å

to the protein). This procedure remains 9,393 pockets (protein–ligand pairs) with 2,707 ligand types.

On this representative pocket database, we benchmarked the retrieval performance of Patch-Surfer using a diverse test set of query pockets that bind either FAD (10), HEM (16), NAD (15), biotin (BTN) (8), fructose 6-phosphate (F6P) (8), guanine (GUN) (10), palmitic acid (PLM) (24), or retinol (RTL) (5) (Fig. 2). In the second parenthesis, the number of query pockets of that type is shown (in total 96 query pockets). For each of these query pockets, pockets in the database were ranked according to the similarity to the query. Then, the retrieval was evaluated in terms of the enrichment factor (EF), which describes the ratio of correctly retrieved pockets relative to the percentage of the database scanned [68, 69]:

$$\mathrm{EF}^{x} = \left(N_{\mathrm{P}}/N_{x}\right) \Big/ \left(T_{\mathrm{P}}/T_{\mathrm{DB}}\right), \quad (1)$$

where T_{P} is the total number of pockets that bind the same ligand type P as the query in the database, T_{DB} is the size of the database, N_{P} is the number of pocket for the ligand type P ranked within the top $X\%$ by the database search method (Patch-Surfer) and N_x is the total number of retrieved pockets ranked in the top $X\%$ of the database. EF is a commonly used metric for evaluating the database retrieval for a large database, for example, in evaluation of methods for drug database search in the cheminformatics domain.

The results are shown in Fig. 2a. At 0.1% retrieval (i.e. considering top 9 pockets), all of the ligands except for two smallest ligand types, F6P and GUN, have high EF values, ranging from around 16.31 (NAD) to over 84.84 (RTL). At 1% retrieval, all of the ligand has EF over 5.0, from 5.13 (F6P) to 49.26 (RTL). The search against this large database was, on average 5–6 min for a query. Having a high EF at a low percentage of retrieval, as we achieved here, is crucial when further computational or experimental validation of binding ligands are to be performed. It is entirely feasible to perform around 100 computational ligand protein docking with consideration of full ligand flexibility [70] or experimental ligand screening in a realistic time. Together with the other types of structure-based function methods, Patch-Surfer as well as Pocket-Surfer will be valuable tools for elucidating function of proteins whose structure are solved by structural genomics efforts. We are currently in the process of making Patch-Surfer available for academic users [66] as a new component of the protein surface comparison server http://kiharalab.org/3d-surfer/ [34]. Users will be able to submit a protein structure, from which pockets regions will be identified and compared against the above-mentioned representative known ligand binding pockets.

Sequence-based function prediction methods that use weakly similar sequences

Sequence-based function prediction methods are applicable for a larger number of proteins than structure-based methods. This is obviously because sequence information is available for the majority of proteins and also because most of function information is stored in sequence databases.

As mentioned in Introduction, conventional methods that are based on homology [18–20] or high sequence-conservation (domains, motifs) [21, 71–73], cover only a small portion of proteins in a genome in terms of function annotation [22, 74]. In recent years, to meet the need of assisting systems biology approaches that deals with a large number of proteins, several novel sequence-based methods have been developed that employ not only highly similar but also weakly similar sequences as the source of function information. The development of such new generation sequence-based approaches is supported by the realization that weakly similar sequences still share functional similarity in many cases [75–77] especially when a

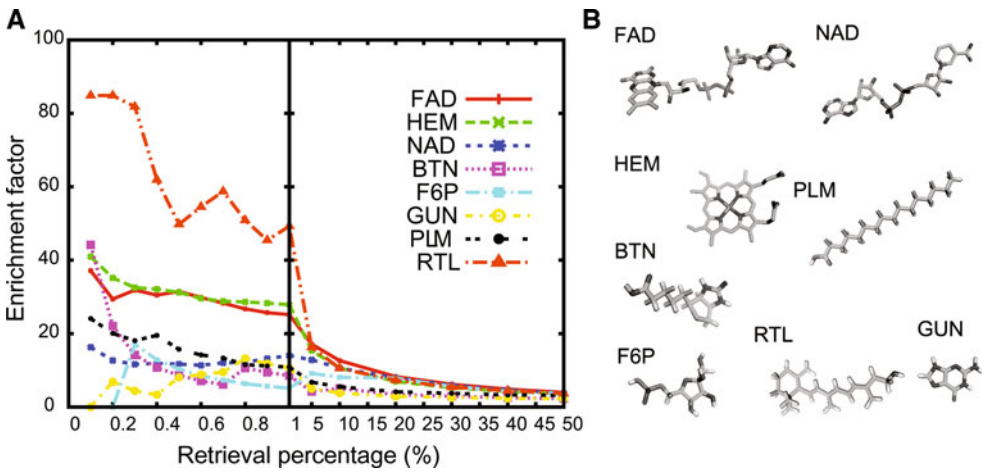

Fig. 2 The enrichment factor for eight types of ligand molecules using Patch-Surfer scanned against the representative binding pocket database. In total 96 query pockets from eight ligand types were used. **a** EF is shown relative to the percentage of top ranking pockets. The EF is averaged for the same types of ligand molecules. **b** The structures of eight ligands used as queries

general level functional category is concerned [78, 79]. Such methods include those which use BLAST or PSI-BLAST search results systematically by applying algorithmic techniques and making use of the Gene Ontology (GO) vocabulary structure [80] (e.g. Gotcha [81], GoFigure [82], OntoBlast [83], PFP [78, 79, 84, 85], ESG [86], and ConFunc [87]). Another direction of recent development is to consider phylogenetic trees aiming more specific function prediction among protein subfamilies (e.g. SIFTER [88], and FlowerPower [89]). JAFA is a meta-server which combines predictions from different servers [90]. A list of more methods for sequence-based function prediction can be found in our recent articles [22, 23, 91].

In the later section we overview two sequence based function prediction methods, protein function prediction (PFP) [78, 79, 84] and extended similarity group (ESG) [86] developed in our group as the examples of these recent methods. PFP makes use of strongly as well as weakly similar sequences to the query sequence and shows improved sensitivity and coverage, whereas ESG draws consensus from the multiple level neighborhoods of similar sequences to improve the precision of predicted GO annotations. Examples of function predictions by PFP using weakly similar sequences are provided.

The protein function prediction algorithm

The PFP algorithm extracts function information (GO terms) from sequences retrieved by PSI-BLAST including very weakly similar sequences with an *E*-value of up to 100. This enables it to predict low resolution terms when there are no homologous sequences available in the database. GO terms are ranked by a raw score computed using Eq. 2. The score for GO term f_a is defined as

$$s(f_a) = \sum_{i=1}^{N} \sum_{j=1}^{Nfunc(i)} \left((-\log(Evalue(i)) + b) P(f_a|f_j) \right) \qquad (2)$$

where N is the number of PSI-BLAST hits obtained for a query sequence, *Nfunc(i)* is the number of GO annotations for the sequence hit i, *E-value(i)* is the PSI-BLAST *E*-value for the sequence hit i, f_j is the jth annotation of the sequence hit i, and constant b takes value 2 (= $\log_{10}100$) to keep the score positive. The conditional probability $P(f_a|f_j)$ indicates the likelihood of having function f_a as an annotation for the query sequence given that f_j is used to annotate the sequence (function association). This function association is computed as the ratio of co-occurrences of terms f_a and f_j in annotations of the same proteins in the UniProt sequence database [92] relative to the number of times term f_j is used to annotate proteins. Since the score for a GO term is basically the sum of weights, $-log(E\text{-}value)$, of all sequences up to very weakly similar ones, consensus annotations from the weakly similar sequence hits can have a high score even if the annotations do not exist in the top sequence hits. Figure 3 illustrates this computation of the raw score for a GO term f_a.

Along with this raw score computation, PFP also transfers the scores of a GO term partially to its less specific parent GO terms in the GO hierarchy in proportion to the ratio of the number of genes annotated by the child and the parent terms. The raw scores are then converted into *p* values using the background score distributions for each term and are further translated into an expected accuracy based on the benchmark dataset.

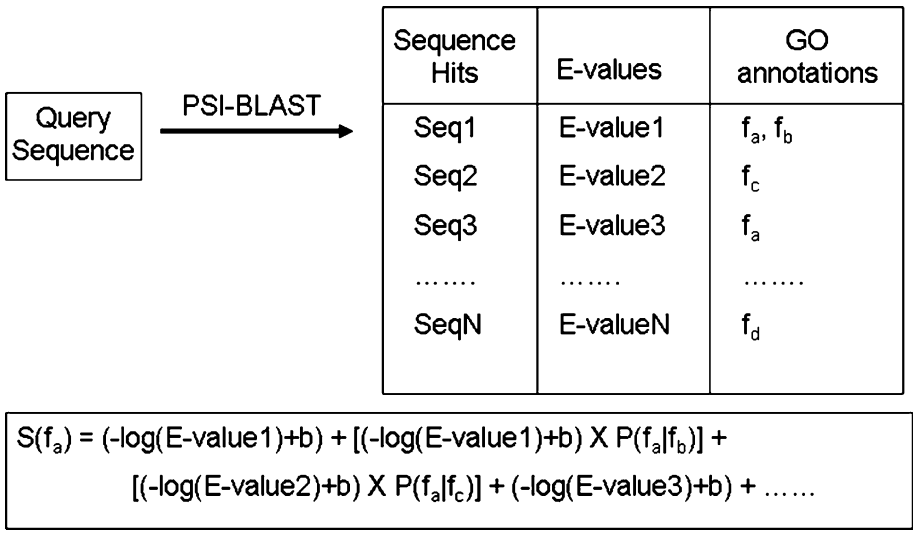

Fig. 3 An example of raw score computation for GO term f_a by the PFP algorithm

The extended similarity group algorithm

Extended similarity group (ESG) runs PSI-BLAST iteratively and advocates functional terms that occur consistently in the series of PSI-BLAST database searches. The algorithm is illustrated in Fig. 4. Starting from the query sequence Q, we first obtain N sequence hits by a PSI-BLAST search, S_1, $S_2 \ldots S_N$, which have E-values E_1, $E_2 \ldots E_N$, respectively. Each sequence hit S_i at the first level is assigned a weight W_i given by Eq. 3, which consists of a normalized E-value of S_i with respect to E-values of all other sequence hits. Then, further from each of the retrieved sequences S_i, PSI-BLAST is run again to retrieve a set of sequence hits, S_{ij}. The weight W_{ij} for each second level sequence is computed similarly as the weight W_i assigned to the first level sequences.

$$Wi = \frac{-\log(E_i) + b}{\sum_{j=1}^{N} \{-\log(E_j) + b\}} \quad (3)$$

Using the weights assigned to sequences retrieved in the first level and the second level searches, the score of a GO term f_a for the query Q is computed as the sum of the weights of sequences which have function f_a:

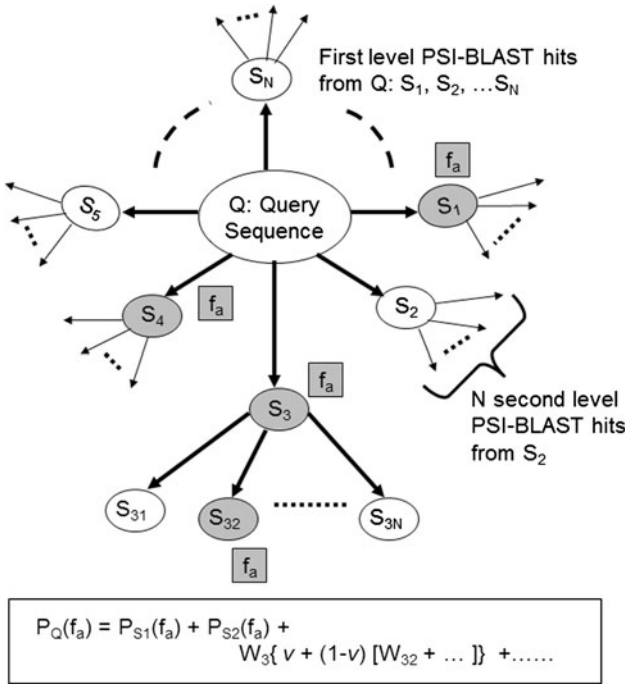

Fig. 4 Computing the probability score for GO term f_a to annotate the query sequence Q using two levels of ESG. In the initial run of PSI-BLAST, N sequences are retrieved. Among them, S_1, S_3, and S_4 are annotated with f_a (colored in *gray*). From each of the retrieved sequences S_1 to S_N, PSI-BLAST is run again, retrieving the second level hits for each of them. S_{32}, a sequence retrieved from S_3, has annotation f_a. The overall score for f_a is the sum of the weights for, S_1, S_3, S_4 and S_{32}

$$P_Q(f_a) = \sum_{i=1}^{N} W_i \cdot P_{Si}(f_a) \quad (4)$$

$$P_{si}(f_a) = v \cdot Is_i(f_a) + (1 - v) \cdot \sum_{j=1}^{N_i} W_{ij} \cdot I_{Sij}(f_a) \quad (5)$$

Equation 4 shows that the score of the GO term f_a for a query sequence Q (the probability that Q has the GO term f_a) is the weighted sum of $P_{si}(f_a)$, the score for f_a assigned to each sequences retrieved in the first level using the Eq. 5. Now Eq. 5 shows that $P_{si}(f_a)$ is the sum of the score $I_{Si}(f_a)$, which is 1 when sequence S_i is annotated with f_a and 0 otherwise, and the weighted sum of the scores that come from sequences retrieved by the second level search for sequence i. The weighting factor v controls contributions from sequences retrieved in the first level and those found in the second level search.

Function prediction using PFP and ESG

PFP and ESG have been thoroughly benchmarked on several datasets including a large one with 11 complete genomes [78, 79, 86]. The benchmark studies for PFP demonstrate its ability to make correct function predictions even in the cases where the query sequence only has hits with large E-values (i.e. insignificant E-values), e.g. E-value of 10 or more in a PSI-BLAST search [78, 79]. By making use of weakly similar sequence hits, PFP can significantly increase annotation coverage of a genome. When PFP was applied to 15 genome sequences, including microbial genomes, *Caenorhabditis elegans*, mouse, *Arabidopsis*, and human genomes, more than two-thirds of the previously unknown proteins in each genome could be assigned a GO function term at the highest confidence level [79]. Predicted function derived mainly from only weakly similar sequence hits are often of low resolution, i.e. GO terms indicating somewhat general function categories that locate at the shallower levels in the GO hierarchy. However, these low resolution functions will be still useful for guiding further detailed investigation of protein function.

To illustrate PFPs' ability to make correct predictions out of weakly similar sequence hits, we showed in Table 1 four examples of PFP's predictions which were computed only from sequences with an E-value above 1.0 or 10.0. Note that smaller E-value indicates more statistically significant hits, and the commonly used E-value cutoff is 0.01 or 0.001. This is to simulate the situation that there are no significant sequence hits in the PSI-BLAST search. The first example is function prediction made for the sequence of outward rectifying potassium channel protein TREK-1 (UniProt ID: O95069). PFP predicts *inward rectifier*

Table 1 Examples of correct annotations predicted by PFP using weakly sequence hits

UniProt ID	GO Annotations	Definition of the GO terms	Relevant PFP predictions using *E*-values > 1.0	Definition of the Predicted GO terms	Rank	Relevant PFP predictions using *E*-values > 10.0	Definition of the Predicted GO terms	Rank
O95069 Outward rectifying potassium channel protein TREK-1	GO:0006813	Potassium ion transport	GO:0004878	Complement component C5a receptor activity	1	GO:0004987	Kappa-opioid receptor activity	1
	GO:0007186	G-protein coupled receptor protein signaling pathway	GO:0001847	Opsonin receptor activity	2	GO:0005242	Inward rectifier potassium channel activity	6
	GO:0071805	Potassium ion transmembrane transport	GO:0005242	Inward rectifier potassium channel activity	3	GO:0001518	Voltage-gated sodium channel complex	2
	GO:0034765	Regulation of ion transmembrane transport	GO:0004987	Kappa-opioid receptor activity	4	GO:0019866	Inner membrane	3
	GO:0005249	Voltage-gated potassium channel activity	GO:0001518	Voltage-gated sodium channel complex	1	GO:0016020	Membrane	4
	GO:0015271	Outward rectifier potassium channel activity	GO:0017071	Intracellular cyclic nucleotide activated cation channel complex	4			
	GO:0016020	Membrane	GO:0019866	Inner membrane	7			
	GO:0016021	Integral to membrane	GO:0016020	Membrane	8			
			GO:0008076	Voltage-gated potassium channel complex	11			
E1WAA4 Formate hydrogenlyase transcriptional activator	GO:0000160	Two-component signal transduction system (phosphorelay)	GO:0005488	Binding	3	GO:0005488	Binding	2
	GO:0006351	Transcription, DNA-dependent	GO:0016462	Pyrophosphatase activity	9	GO:0016462	Pyrophosphatase activity	9
	GO:0006355	Regulation of transcription, DNA-dependent	GO:0003677	DNA binding	11	GO:0003677	DNA binding	10
	GO:0000166	Nucleotide binding	GO:0003700	Transcription factor activity	14	GO:0006351	Transcription, DNA-dependent	4
	GO:0003677	DNA binding	GO:0050907	Sensory transduction of chemical stimulus	6			
	GO:0003700	Sequence-specific DNA binding transcription factor activity	GO:0006351	Transcription, DNA-dependent	7			
	GO:0005524	ATP binding						
	GO:0008134	Transcription factor binding						

Table 1 continued

UniProt ID	GO Annotations	Definition of the GO terms	Relevant PFP predictions using *E*-values > 1.0	Definition of the Predicted GO terms	Rank	Relevant PFP predictions using *E*-values > 10.0	Definition of the Predicted GO terms	Rank
	GO:0017111	Nucleoside-triphosphatase activity						
Q8UVE6 Transcription factor AP2 alpha 1	GO:0001501	Skeletal system development	GO:0003705	RNA polymerase II transcription factor activity, enhancer binding	6	GO:0008134	Transcription factor binding	7
	GO:0006351	Transcription, DNA-dependent	GO:0003677	DNA binding	7	GO:0003700	Transcription factor activity	8
	GO:0006355	Regulation of transcription, DNA-dependent	GO:0003700	Transcription factor activity	13	GO:0006351	Transcription, DNA-dependent	2
	GO:0007422	Peripheral nervous system development	GO:0008134	Transcription factor binding	14			
	GO:0014036	Neural crest cell fate specification	GO:0006351	Transcription, DNA-dependent	3			
	GO:0030318	Melanocyte differentiation						
	GO:0060041	Retina development in camera-type eye						
	GO:0003700	Sequence-specific DNA binding transcription factor activity						
	GO:0005634	Nucleus						
Q12386 Actin-like protein ARP8	GO:0006312	Mitotic recombination	GO:0005488	Binding	1	GO:0003676	Nucleic acid binding	4
	GO:0006338	Chromatin remodeling	GO:0003676	Nucleic acid binding	4	GO:0003682	Chromatin binding	7
	GO:0006974	Response to DNA damage stimulus	GO:0003682	Chromatin binding	7	GO:0008135	Translation factor activity, nucleic acid binding	13
	GO:0006355	Regulation of transcription, DNA-dependent	GO:0003677	DNA binding	12	GO:0003723	RNA binding	16
	GO:0003729	mRNA binding	GO:0003697	Single-stranded DNA binding	13	GO:0046034	ATP metabolism	3
	GO:0043140	ATP-dependent 3′-5′ DNA helicase activity	GO:0003700	Transcription factor activity	15	GO:0009199	Ribonucleoside triphosphate metabolism	9
	GO:0005634	Nucleus	GO:0006351	Transcription, DNA-dependent	2	GO:0006351	Transcription, DNA-dependent	13
	GO:0005856	Cytoskeleton	GO:0046034	ATP metabolism	6			

Table 1 continued

UniProt ID	GO Annotations	Definition of the GO terms	Relevant PFP predictions using *E*-values > 1.0	Definition of the Predicted GO terms	Rank	Relevant PFP predictions using *E*-values > 10.0	Definition of the Predicted GO terms	Rank
	GO:0031011	Ino80 complex	GO:0009199	Ribonucleoside triphosphate metabolism	9			

Function predictions by PFP for four proteins are shown. Annotations from the UniProt database are shown in the first three columns from left. The next three columns (from the 4th to the 6th column) show prediction by PFP that are derived only from weak sequence hits with an *E*-value of 1.0 or larger. Only the predictions relevant to the correct annotations are shown. "Rank" is the rank of the prediction based on the PFP's confidence score. Since the predicted GO terms are ranked for each of the three GO categories separately, there are multiple (up to 3) predictions with the same rank. The last three columns (the 7th–9th column) are predictions by PFP using weak sequence hits with an *E*-value of 10.0 or larger

potassium channel activity (*GO:0005242)* with *E*-value cutoffs of both 1.0 and 10.0. Although this prediction does not exactly match with this protein's annotation, it is close in the GO hierarchy to the correct annotation, *outward rectifier potassium channel activity (GO:0015271)*. Both terms have a common immediate parent terms, *voltage-gated potassium channel activity (GO:0005249)*. This query protein is involved in the *G-protein coupled receptor protein signaling pathway (GO:0007186)*, for which PFP using the *E*-value cutoff of 1.0 and 10.0 has captured more specialized child terms of *GO:0004888 trans membrane signaling receptor activity* and *GO:0004930 G-protein coupled receptor activity* (e.g. *kappa-opioid receptor activity*, GO: 0004987). Overall in this example, even using weak sequence hits, which are conventionally discarded in the homology search, PFP still managed to indicate that this protein is potassium channel that locate inner membrane (transmembrane).

The second example of formate hydrogenlyase transcriptional activator (Uniprot ID: E1WAA4) is involved in *transcription, DNA-dependent (GO:0006351)*. This GO term was predicted by PFP within the top 10 ranks when using the *E*-value cutoff of 1.0 and 10.0. Also this protein is annotated with *GO:0017111 nucleoside-triphosphatase activity*, where PFP predicts a less specific parental term, *GO:0016462 pyrophosphatase activity* as an annotation. Similar results can be seen in the last two examples for Q8UVE6 and Q12386. Using only sequence hits of *E*-value above 1.0/10.0, PFP correctly predicted their functional class, transcription factor. More examples can be found in the original paper [79].

PFP's superior performance has been also demonstrated in the community-wide computational function prediction assessments. In Automatic Function Prediction Special Interest Group (AFP-SIG) meeting held at the Intelligent System in Molecular Biology (ISMB) AFP-SIG 2005 [93] and the function prediction category at the Critical Assessment of Techniques for Protein Structure Prediction 7 (CASP7) [94], PFP has shown best overall performance among the participants.

In contrast to PFP whose aim is to increase the sensitivity to enlarge annotation coverage, ESG is intended to make more precise prediction by iterative database searches. In the thorough benchmark study [86], ESG was found to have a higher precision than PFP and the other existing methods with a comparable sensitivity to PFP. ESG was found to have more accurate prediction for multi-domain proteins since the second round of PSI-BLAST searches are often initiated from different local regions of the query sequence.

Availability of PFP and ESG

PFP and ESG are available freely for academic users as web servers at http://kiharalab.org/web/pfp.php and http://kiharalab.org/web/esg.php. The users can submit sequences and receive predicted GO terms for the sequences. The stand-alone PFP program is available upon request.

Conclusion

Many protein structures determined by the structural genomics projects remain functionally unknown since they are not homologous to or do not have the global sequence or structural similarity to characterized proteins. In this article, we have discussed structure-based methods and sequence-based methods developed in our group to cope with such proteins with unknown function. Two structure-based methods, Pocket-Surfer and Patch-Surfer, detect similar known binding pockets for pocket regions in a query protein without using global protein fold similarity. Two sequence-based methods, PFP and ESG, make use of weakly similar sequences that are conventionally discarded in homology based function annotation. Combined together

with experimental methods we hope that computational methods will make a leading contribution in functional elucidation of the protein structures.

Acknowledgments This work is supported in part by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM075004, R01GM097528), the National Science Foundation (DMS0800568, EF0850009, IIS0915801) and Showalter Trust. MC is supported by Bilsland Dissertation Fellowship from College of Science, Purdue University.

References

1. Chandonia JM, Brenner SE (2006) The impact of structural genomics: expectations and outcomes. Science 311:347–351
2. Norvell JC, Berg JM (2007) Update on the protein structure initiative. Structure 15:1519–1522
3. Terwilliger TC, Stuart D, Yokoyama S (2009) Lessons from structural genomics. Annu Rev Biophys 38:371–383
4. Todd AE, Marsden RL, Thornton JM, Orengo CA (2005) Progress of structural genomics initiatives: an analysis of solved target structures. J Mol Biol 348:1235–1260
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242
6. Westbrook J, Feng Z, Chen L, Yang H, Berman HM (2003) The protein data bank and structural genomics. Nucleic Acids Res 31:489–491
7. Ellrott K, Zmasek CM, Weekes D, Sri KS, Bakolitsa C, Godzik A, Wooley J (2011) TOPSAN: a dynamic web database for structural genomics. Nucleic Acids Res 39:D494–D496
8. Shin DH, Hou J, Chandonia JM, Das D, Choi IG, Kim R, Kim SH (2007) Structure-based inference of molecular functions of proteins of unknown function from Berkeley Structural Genomics Center. J Struct Funct Genomics 8:99–105
9. Teplyakov A, Pullalarevu S, Obmolova G, Doseeva V, Galkin A, Herzberg O, Dauter M, Dauter Z, Gilliland GL (2004) Crystal structure of the YffB protein from Pseudomonas aeruginosa suggests a glutathione-dependent thiol reductase function. BMC Struct Biol 4:5
10. Teplyakov A, Obmolova G, Sarikaya E, Pullalarevu S, Krajewski W, Galkin A, Howard AJ, Herzberg O, Gilliland GL (2004) Crystal structure of the YgfZ protein from *Escherichia coli* suggests a folate-dependent regulatory role in one-carbon metabolism. J Bacteriol 186:7134–7140
11. Li De La Sierra-Gallay I, Collinet B, Graille M, Quevillon-Cheruel S, Liger D, Minard P, Blondeau K, Henckes G, Aufrere R, Leulliot N, Zhou CZ, Sorel I, Ferrer JL, Poupon A, Janin J, van Tilbeurgh H (2004) Crystal structure of the YGR205w protein from Saccharomyces cerevisiae: close structural resemblance to *E. coli* pantothenate kinase. Proteins 54:776–783
12. Graille M, Quevillon-Cheruel S, Leulliot N, Zhou CZ, Gallay ILD, Jacquamet L, Ferrer JL, Liger D, Poupon A, Janin J, van Tilbeurgh H (2004) Crystal structure of the YDR533c S. cerevisiae protein, a class II member of the Hsp31 family. Structure 12:839–847
13. Liger D, Graille M, Zhou CZ, Leulliot N, Quevillon-Cheruel S, Blondeau K, Janin J, van Tilbeurgh T (2004) Crystal structure and functional characterization of yeast YLR011wp, an enzyme with NAD(P)H-FMN and ferric iron reductase activities. J Biol Chem 279:34890–34897
14. Sanishvili R, Yakunin AF, Laskowski RA, Skarina T, Evdokimova E, Doherty-Kirby A, Lajoie GA, Thornton JM, Arrowsmith CH, Savchenko A, Joachimiak A, Edwards AM (2003) Integrating structure, bioinformatics, and enzymology to discover function: BioH, a new carboxylesterase from *Escherichia coli*. J Biol Chem 278:26039–26045
15. Kuznetsova E, Proudfoot M, Sanders SA, Reinking J, Savchenko A, Arrowsmith CH, Edwards AM, Yakunin AF (2005) Enzyme genomics: application of general enzymatic screens to discover new enzymes. FEMS Microbiol Rev 29:263–279
16. Fridman E, Pichersky E (2005) Metabolomics, genomics, proteomics, and the identification of enzymes and their substrates and products. Curr Opin Plant Biol 8:242–248
17. Roberts RJ (2011) COMBREX: COMputational BRidge to EXperiments. Biochem Soc Trans 39:581–583
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
19. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
20. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448
21. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. Nucleic Acids Res 38:D211–D222
22. Hawkins T, Kihara D (2007) Function prediction of uncharacterized proteins. J Bioinform Comput Biol 5:1–30
23. Hawkins T, Chitale M, Kihara D (2008) New paradigm in protein function prediction for large scale omics analysis. Mol Biosyst 4:223–231
24. Kihara D (2011) Protein function prediction for omics era. Springer, London
25. Gherardini PF, Helmer-Citterich M (2008) Structure-based function prediction: approaches and applications. Brief Funct Genomic Proteomic 7:291–302
26. Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM (1998) Protein folds and functions. Structure 6:875–884
27. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA (2000) From structure to function: approaches and limitations. Nat Struct Biol 7(Suppl):991–994
28. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11:739–747
29. Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. J Mol Biol 233:123–138
30. Orengo CA, Taylor WR (1996) SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol 266:617–635
31. Thompson KE, Wang Y, Madej T, Bryant SH (2009) Improving protein structure similarity searches using domain boundaries based on conserved sequence information. BMC Struct Biol 9:33
32. Mizuguchi K, Go N (1995) Comparison of spatial arrangements of secondary structural elements in proteins. Protein Eng 8: 353–362
33. Kihara D, Sael L, Chikhi R, Esquivel-Rodriguez J (2011) Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking. Curr Protein Pept Sci 12:520–530
34. La D, Esquivel-Rodriguez J, Venkatraman V, Li B, Sael L, Ueng S, Ahrendt S, Kihara D (2009) 3D-SURFER: software for high-throughput protein surface comparison and analysis. Bioinformatics 25:2843–2844
35. Sael L, Li B, La D, Fang Y, Ramani K, Rustamov R, Kihara D (2008) Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins 72:1259–1273

36. Sael L, Kihara D (2009) Protein surface representation and comparison: new approaches in structural proteomics. In: Chen J, Lonardi S (eds) Biological data mining. Chapman & Hall/CRC Press, Boca Raton, pp 89–109
37. Venkatraman V, Sael L, Kihara D (2009) Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors. Cell Biochem Biophys 54:23–32
38. Ritchie DW, Graham J (1999) Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. J Comp Chem 20:383–395
39. Orengo CA, Jones DT, Thornton JM (1994) Protein superfamilies and domain superfolds. Nature 372:631–634
40. Porter CT, Bartlett GJ, Thornton JM (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32:D129–D133
41. Arakaki AK, Zhang Y, Skolnick J (2004) Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. Bioinformatics 20:1087–1096
42. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P (1994) A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol 243:327–344
43. Kleywegt GJ (1999) Recognition of spatial motifs in protein structures. J Mol Biol 285:1887–1897
44. Ferre F, Ausiello G, Zanzoni A, Helmer-Citterich M (2004) SURFACE: a database of protein surface regions for functional annotation. Nucleic Acids Res 32:D240–D244
45. Redfern OC, Dessailly BH, Dallman TJ, Sillitoe I, Orengo CA (2009) FLORA: a novel method to predict protein function from structure in diverse superfamilies. PLoS Comput Biol 5:e1000485
46. Schmitt S, Kuhn D, Klebe G (2002) A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol 323:387–406
47. Gold ND, Jackson RM (2006) Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. J Mol Biol 355:1112–1124
48. Kinoshita K, Nakamura H (2005) Identification of the ligand binding sites on the molecular surface of proteins. Protein Sci 14:711–718
49. Morris RJ, Najmanovich RJ, Kahraman A, Thornton JM (2005) Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. Bioinformatics 21:2347–2355
50. Binkowski TA, Adamian L, Liang J (2003) Inferring functional relationships of proteins from local sequence and spatial surface patterns. J Mol Biol 332:505–526
51. Binkowski TA, Freeman P, Liang J (2004) pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. Nucleic Acids Res 32:W555–W558
52. Binkowski TA, Joachimiak A (2008) Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. BMC Struct Biol 8:45
53. Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33:W89–W93
54. Pal D, Eisenberg D (2005) Inference of protein function from protein structure. Structure (Camb) 13:121–130
55. Chikhi R, Sael L, Kihara D (2010) Real-time ligand binding pocket database search using local surface descriptors. Proteins 78:2007–2028
56. Sael L, Kihara D (2011) Binding ligand prediction for proteins using partial matching of local surface patches. Int J Mol Sci 11:5009–5026
57. Sael L, Kihara D (2012) Detecting local ligand-binding site similarity in non-homologous proteins by surface patch comparison. Proteins (in press)
58. Novotni M, Klein R (2003) 3D Zernike descriptors for content based shape retrieval. In: ACM symposium on solid and physical modeling, proceedings of the eighth ACM symposium on solid modeling and applications pp 216–225
59. Canterakis N (1999) 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. In: Proceedings of 11th scandinavian conference on image analysis, pp 85–93
60. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA 98:10037–10041
61. Li B, Turuvekere S, Agrawal M, La D, Ramani K, Kihara D (2007) Characterization of local geometry of protein surfaces with the visibility criterion. Proteins 71:670–683
62. Huang B, Schroeder M (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol 6:19
63. Demange G, Gale D, Stomayor M (1986) Multi-item auctions. J Polit Econ 94:863–872
64. Kahraman A, Morris RJ, Laskowski RA, Favia AD, Thornton JM (2010) On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. Proteins 78:1120–1136
65. Gribskov M, Robinson NL (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput Chem 20:25–33
66. Sael L, Kihara D (2012) Constructing patch-based ligand-binding pocket database for predicting function of proteins. BMC Bioinform (in press)
67. Wallach I, Lilien R (2009) The protein-small-molecule database, a non-redundant structural resource for the analysis of protein-ligand binding. Bioinformatics 25:615–620
68. Bender A, Glen RC (2005) A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. J Chem Inf Model 45:1369–1375
69. Venkatraman V, Chakravarthy PR, Kihara D (2009) Application of 3D Zernike descriptors to shape-based ligand similarity searching. J Cheminform 1:19
70. Huang SY, Zou X (2010) Advances and challenges in protein-ligand docking. Int J Mol Sci 11:3016–3034
71. Hulo N, Bairoch A, Bulliard V, Cerutti L, De CE, Langendijk-Genevaux PS, Pagni M, Sigrist CJ (2006) The PROSITE database. Nucleic Acids Res 34:D227–D230
72. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH (2005) InterPro, progress and status in 2005. Nucleic Acids Res 33:D201–D205
73. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D (2005) The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res 33:D212–D215
74. Chitale M, Kihara D (2011) Computational protein function prediction: framework and challenges. In: Kihara D (ed) Protein function prediction for omis era. Springer, London, pp 1–17
75. John B, Sali A (2004) Detection of homologous proteins by an intermediate sequence search. Protein Sci 13:54–62
76. Salamov AA, Suwa M, Orengo CA, Swindells MB (1999) Combining sensitive database searches with multiple intermediates to detect distant homologues. Protein Eng 12:95–100
77. Park J, Teichmann SA, Hubbard T, Chothia C (1997) Intermediate sequences increase the detection of homology between sequences. J Mol Biol 273:349–354

78. Hawkins T, Luban S, Kihara D (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 15:1550–1556
79. Hawkins T, Chitale M, Luban S, Kihara D (2009) PFP: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins 74: 566–582
80. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–D261
81. Martin DM, Berriman M, Barton GJ (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinform 5:178
82. Khan S, Situ G, Decker K, Schmidt CJ (2003) GoFigure: automated gene ontology annotation. Bioinformatics 19:2484–2485
83. Zehetner G (2003) OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms. Nucleic Acids Res 31:3799–3803
84. Hawkins T, Chitale M, Kihara D (2010) Functional enrichment analyses and construction of functional similarity networks with high confidence function prediction by PFP. BMC Bioinform 11:265
85. Si L, Yu D, Kihara D, Yi F (2008) Combining sequence similarity scores and textual information for gene function annotation in the literature. Inf Retr 11:389–404
86. Chitale M, Hawkins T, Park C, Kihara D (2009) ESG: extended similarity group method for automated protein function prediction. Bioinformatics 25:1739–1745
87. Wass MN, Sternberg MJ (2008) ConFunc–functional annotation in the twilight zone. Bioinformatics 24:798–806
88. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE (2005) Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol 1:e45
89. Krishnamurthy N, Brown D, Sjolander K (2007) FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. BMC Evol Biol 7(Suppl 1):S12
90. Friedberg I, Harder T, Godzik A (2006) JAFA: a protein function annotation meta-server. Nucleic Acids Res 34:W379–W381
91. Chitale M, Hawkins T, Kihara D (2009) Automated prediction of protein function from sequence. In: Bujnicki J (ed) Prediction of protein structure, functions, and interactions. Wiley, London, pp 63–86
92. Uniprot Consortium (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38:D142–D148
93. Friedberg I, Jambon M, Godzik A (2006) New avenues in protein function prediction. Protein Sci 15:1527–1529
94. Lopez G, Rojas A, Tress M, Valencia A (2007) Assessment of predictions submitted for the CASP7 function prediction category. Proteins 69:165–174