Protein Function Annotation Using Protein Domain Family Resources

Sayoni Das and Christine A Orengo

Address
Institute of Structural and Molecular Biology, UCL, Gower Street, WC1E 6BT, UK

Corresponding author: Orengo, Christine A (c.orengo@ucl.ac.uk)

Abstract

As a result of the genome sequencing and structural genomics initiatives, we have a wealth of protein sequence and structural data. However, only about 1% of these proteins have experimental functional annotations. As a result, computational approaches that can predict protein functions are essential in bridging this widening annotation gap. This article reviews the current approaches of protein function prediction using structure and sequence based classification of protein domain family resources with a special focus on functional families in the CATH-Gene3D resource.

Keywords: Protein function prediction, Protein function annotation, Protein family, Protein classification, Moonlighting protein

1. Introduction

Knowledge of the functions of all proteins is key to understanding the nature of the protein universe and in essence, biology. The availability of complete genome sequences and development of high throughput tools for function annotation has been a significant step towards this. The Genomes Online Database [1], which is a centralized resource of genome-sequencing projects worldwide, lists > 64000 sequencing projects as of June 2015, and these are expected to hugely increase the numbers of known sequences in UniProtKB [2]. In contrast, ~1% of the proteins in the current UniProt database (June 2015) are experimentally characterised and it is evident that the current rate of experimental annotations and manual curation process will never be sufficient for complete annotation of the proteins captured in public databases [3]. Therefore, many computational approaches, using both sequence and structural data, have been developed to bridge this widening function annotation gap.

The conventional method used for inferring functional annotations for uncharacterised proteins is a sequence or structure homology search of a query protein against a database of characterised proteins e.g. by BLAST [4] or CATHEDRAL [5] followed by pair-wise annotation

transfer, based on the principle that evolutionarily-related proteins having high sequence or structural similarity have similar, if not identical functions [6]. However, functional inference using such simple similarity metrics [7] can often lead to erroneous functional assignments when sequences diverge (sequence identity < 60%) [6], due to the complex protein function-evolution relationship [8], and in the case of multi-domain [9] and moonlighting proteins [10] or due to any mis-annotations existing in the databases [11].

To address the challenging task of assignment of reliable functions to proteins of unknown function, many recent annotation approaches involve use of protein family resources. Protein family resources cluster protein sequences into families and subfamilies based on their sequence, structure or function similarity (in the case of annotated protein sequences).

2. Protein Family Resources

Classification or clustering of the known parts of the protein universe into homologous groups, has become a popular approach for providing valuable insights into our understanding of the protein function repertoire and how it evolves. In recent years, it has been observed that homologous proteins can often evolve different functions as a result of different sets of residues in their active site [12], addition of secondary structure embellishments to the core protein structure which alters the geometry of the active site of the protein or an interface on the protein [13] or due to domain-shuffling in multi-domain proteins [14] which can alter the context of the domain and again result in changes to functional sites. The identification of protein families and characterization of their functional sites is of utmost importance in understanding how function is modulated during evolution by sequence and structural changes in diverse families [15]. Moreover, understanding the evolution of function in proteins also provides invaluable information that can be useful in protein engineering for designing protein scaffolds with novel functions [16].

Because of the significant divergence of function between relatives in many of the universal and highly populated protein families, one of the major challenges of using these resources for functional annotation is the sub-classification of relatives in these families into coherent functional groups. As well as increasing the accuracy of functional inheritance between relatives, such functional grouping would also facilitate multiple sequence alignment of the relatives to find conserved residue positions which can provide valuable insights about the key functional sites and mechanisms of the protein.

2.1 Whole Protein Families

There are a number of high-quality protein family resources like PANTHER [17], TIGRFAMs [18] and HAMAP [19] among others, which provide manually-curated functional clusters of protein

sequences. However, they are limited by low sequence coverage (Table 1). Using automated approaches, PhyloFacts [20], a phylogenomic encyclopedia of protein families across the Tree of Life, classifies its families into subfamilies using the SCI-PHY algorithm [21] which uses only sequence information. The SCI-PHY (Subfamily Classification in Phylogenomics) algorithm exploits Bayesian and information-theoretic measures to construct a hierarchical phylogenetic tree and define an optimal cut of the tree into subfamilies [22]. Secator [23] is another phylogenomic subfamily identification method which uses a sequence dissimilarity measure in order to cut a phylogenetic tree. These methods invariably require an accurate multiple sequence alignment of the protein family as a starting point in their pipeline which is likely to be erroneous for very large and very diverse families. ProtoNet [24] provides an automatic classification of similar proteins which are further sub-classified into clusters using an information-theoretic protocol [25] based on available annotations. Other subfamily identification methods are available which define clusters using pairwise similarity e.g. CluSTr [26], COGs [27], OrthoMCL [28] and eggNOG [29]. CluSTr, similar to ProtoNet, clusters protein sequences into a hierarchical tree of clusters while OrthoMCL and eggNOG increases the functional accuracy of clustering by restricting to orthologs.

Use of whole protein family resources for function annotation

Protein family resources may be exploited for annotating uncharacterized sequences by mapping query sequences to the best matched family and inheriting the annotations from the characterised sequences. Manually-curated Gene Ontology (GO) [30] term associations are readily available from certain family resources such as TIGRFAM (TIGRFAM2GO) and HAMAP (HAMAP2GO). BAR+ [31], an automated annotation method based on the annotation transfer from protein families, produces clusters such that the pairwise sequence identity between relatives in a cluster is 40% with at least 90% of sequences in the pairwise alignment overlapping. A BLAST search of query sequences against the BAR+ clusters is performed and statistically validated GO annotations are then inferred for the sequences based significant sequence identity and coverage of the match.

Resource	Whole Protein/ Domain	Classification type	Sequence Coverage	Refs.
PANTHER	Whole protein	Curated	1,424,953 genes (v9.0)	[17]
TIGRFAMs	Whole protein	Curated	>58,000 proteins (v15)	[18]

HAMAP	Whole protein	Curated	~ 1 0 , 8 7 4 , 3 5 proteins (as of Sept. 2014)	[19]
PhyloFacts	Whole protein and Domain	Automated	> 7 , 3 0 0 , 0 0 0 proteins (v3.0)	[20]
CluSTr	Whole protein	Automated	15,767,981 proteins (2014)	[26]
COGs	Whole protein	Semi-automated	Sequences from 711 genomes	[27]
OrthoMCL	Whole protein	Automated	1,398,546 proteins from 150 genomes (release 5)	[28]
eggNOG	Whole protein	Automated	4,396,591 proteins (v3.0)	[29]
Pfam	Domain	Curated	~ 1 8 , 8 0 0 , 0 0 0 (v27.0)	[32]
S C O P and SUPERFAMILY	Domain	Curated	4 1 , 9 1 6 , 8 2 4 sequences from 3 2 4 5 distinct organisms (v1.75)	[33,34]
CATH-Gene3D	Domain	Automated	2 1 , 6 6 2 , 1 5 5 sequences from 6 1 3 1 genomes (C A T H v 4 . 1 , Gene3D v14)	[35,36]

Table 1: Resources providing classifications of protein families

2.2 Protein domain families

Proteins are generally composed of one or more distinct, compact units of protein structure called domains that form the functional building blocks of proteins. Multi-domain proteins complicate the protein sequence-structure-function relationship further as they expand the functional repertoire [14]. Consequently, when an uncharacterised protein does not match any annotated protein along their entire length and cannot be assigned to any characterised 'whole protein' families, function can perhaps be better understood by analysing the domain components and finding homologs to each domain.

There are many protein domain resources which provide classifications of protein domains based on either sequence (e.g Pfam [31]) or structure (e.g. CATH [35], SCOP [33] and ECOD [37]). PhyloFacts [20] also provides domain families in its resources which are sub-classified using SCI-PHY, as for the whole protein families.

Pfam [31] is a comprehensive database of protein families which currently provides ~ 80%

coverage of the UniprotKB sequence space. The Funshift [38] database, which provides analysis of function shifts of sequences within a Pfam family, further classifies them into subfamilies using the SCI-PHY algorithm [21]. Meta-protein domain resources like InterPro [39] and the Conserved Domain Database (CDD) [40] combine multiple protein domain family databases, providing higher sequence coverage compared to individual resources.

The structure classification databases, CATH [35] and SCOP [33], classify evolutionary related protein domains into superfamilies. SCOP [33] subclassifies its superfamilies into families by expert curation. However, these families have been found to more closely resemble taxonomic groups rather than functional groups. The Gene3D [36] and SUPERFAMILY [34] resources predict domain sequences belonging to the CATH and SCOP structural superfamilies, respectively. This is done using HMM based strategies and for sequences in UniProt these domain annotations are also made available via the InterPro website [39].

Some of the highly-populated CATH-Gene3D superfamilies can be extremely functionally diverse [41]. Some of the divergence between relatives can be attributed to structural embellishments to the core domain structure and changes in the domain composition of the parent proteins (Figure 1a,1b). To address this diversity superfamilies in CATH are sub-classified into functionally coherent groups of relatives or functional families (FunFams). The starting point of the functional sub-classification in CATH-Gene3D is a hierarchical agglomerative clustering algorithm, GeMMA [42]. GeMMA (Genome Modelling and Model Annotation) clusters close homologues (sequences with at least 90% sequence identity) into starting clusters using CD-HIT [43]. Multiple sequence alignments for each starting cluster are built using MAFFT [44]. GeMMA then performs an iterative all-against-all profile-profile comparison of a set of clusters using COMPASS [45] followed by merging of the most similar clusters and realignment of sequences in the merged cluster by MAFFT. This iterative process continues until one cluster remains per CATH superfamily. The merging order is then used to build a tree of clusters (GeMMA tree) from the leaf nodes to the root rode.

The GeMMA tree for a particular superfamily is then used to classify the superfamily into functional families (Figure 1c) by partitioning it in different ways: (i) coarse functional families can be obtained using an unsupervised method which cuts the hierarchical tree at a generic threshold into families, (ii) a more sophisticated approach DFX (Domain Family Exploration) [46], which utilizes available functional annotation data from Gene Ontology (GO) to ensure functional coherence in the resulting families and (iii) FunFHMMer [47], which utilizes evolutionary signals (specificity-determining positions or SDPs and conserved positions) in cluster multiple-sequence alignments (MSAs) to ensure functional coherence in the resulting families. It has been recently shown that functional classification using FunFHMMer provides more functionally coherent families than those generated by DFX and that the functional families correspond well with the manually-curated classification in the Structure-Function Linkage Database (SFLD) [48]. As the FunFams are predicted to be functionally coherent, functionally important residues (e.g. catalytic residues, ligand-binding residues) in the FunFams are expected to be highly conserved across the family

(see Figure 2). A residue-enrichment analysis (see Figure 2a) of the FunFams demonstrated that conserved residues detected in the FunFams are significantly enriched in known catalytic residues (p-value < 3.64E-51) [49]. Conserved residues were identified by running the program Scorecons [50] on the multiple sequence alignment of FunFam relatives.

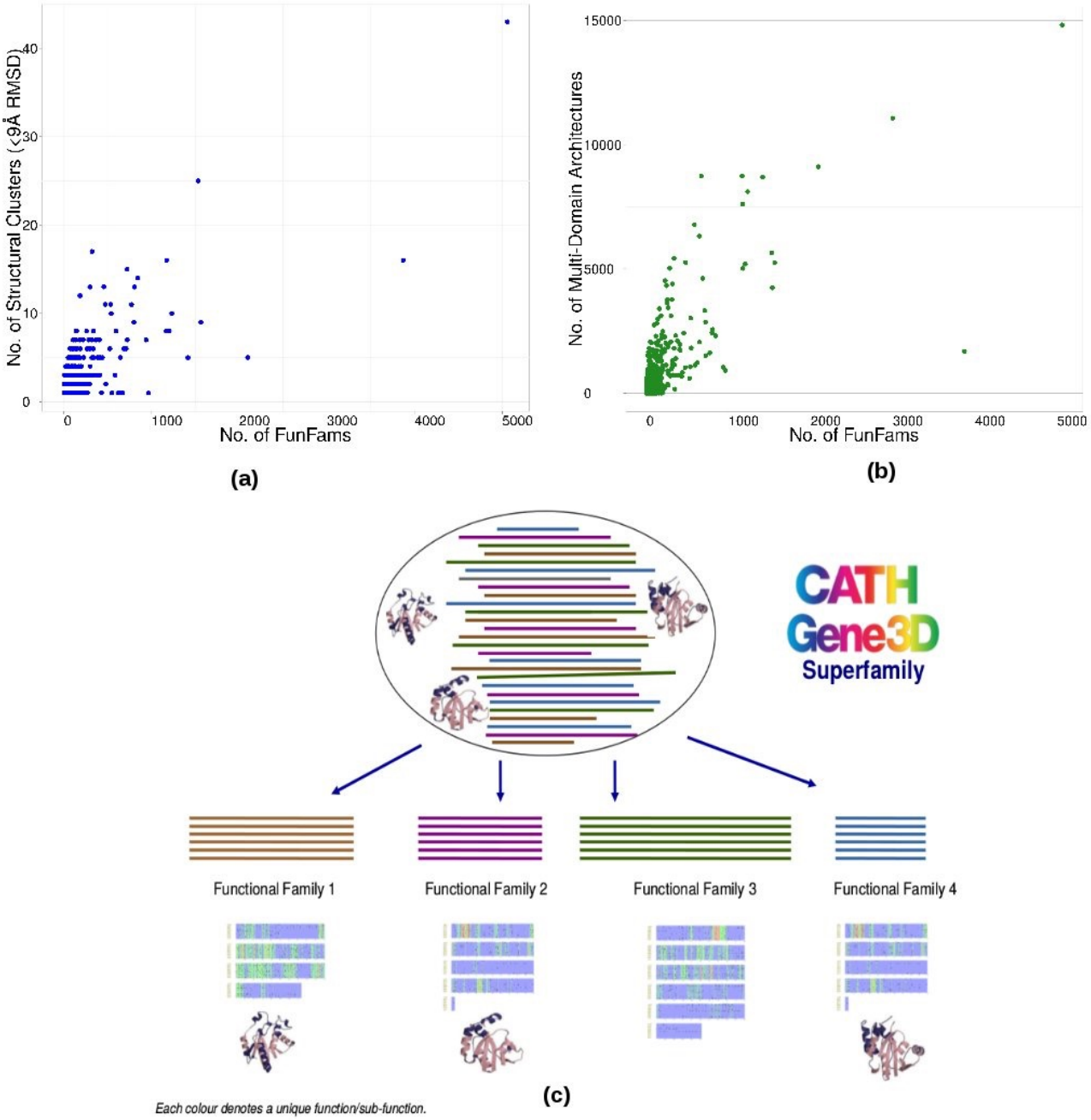

Figure 1: The relationship between the number of functional families in a superfamiliy and the **(a)** structural diversity of the domains in the superfamily (a structural cluster is a group of relatives whose structures can be pair-wise superimposed with an RMSD < 9Å) **(b)** number of different multi-domain architectures (MDAs) in which domain relatives are found. **(c)** Schematic representation of functional sub-classification of domain sequence and structural relatives of a CATH superfamily into functional families (FunFams). Diverse sequence patterns reveal differences in the highly conserved residues in the different FunFams, reflecting differences in the functional properties of the FunFams.

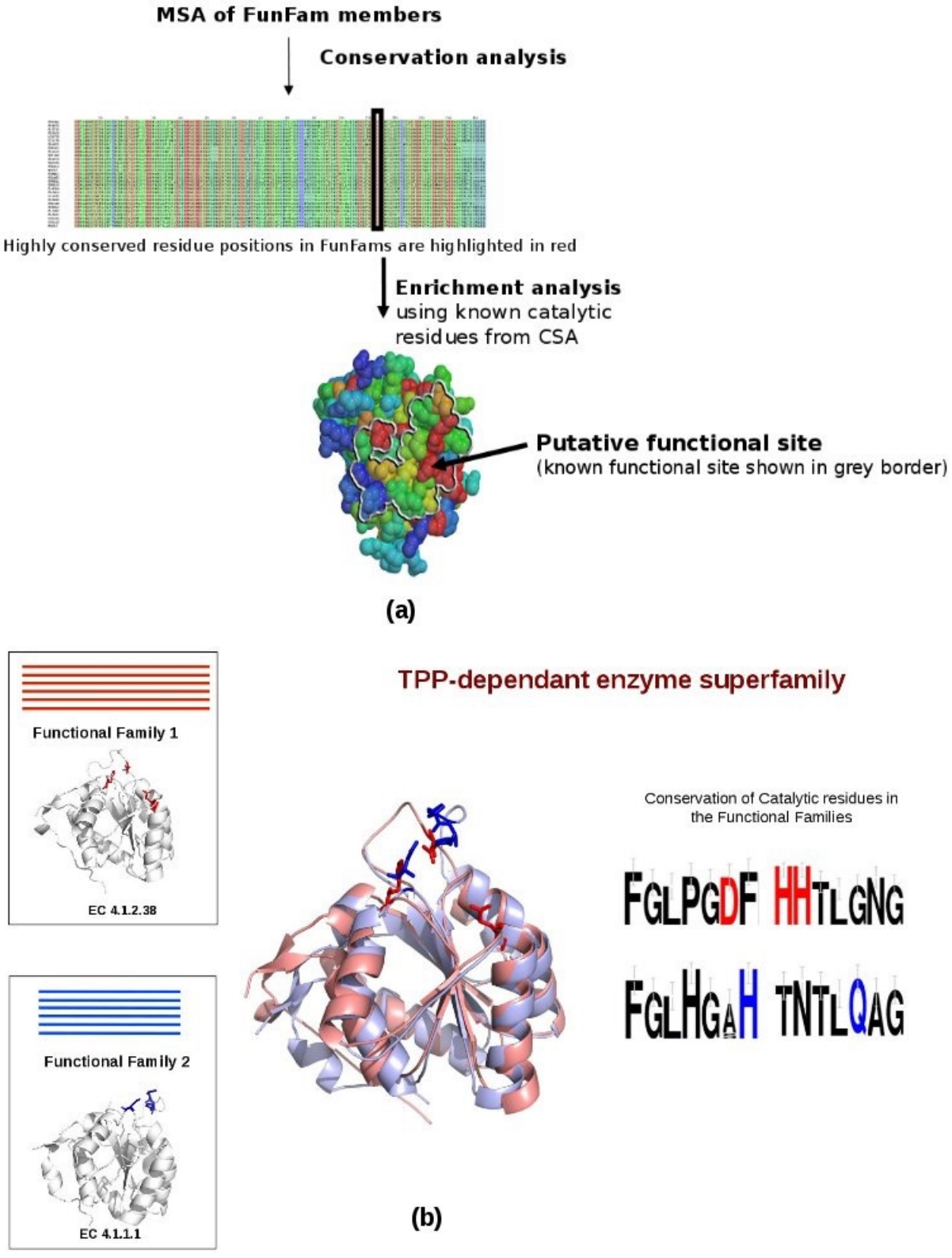

Figure 2: (a) Protocol for the residue enrichment analysis of FunFam alignments. The Scorecons [50] method was used to detect highly conserved residues in the FunFam and these highly conserved residues were found to be significantly enriched in known catalytic residues (p-value < 3.64E-51). **(b)** Differences in the catalytic residues between two FunFams in the Thiamine diphosphate-dependant enzyme superfamily, having different EC numbers. The catalytic residues for domains belonging to Functional Families 1 and 2 are shown in red and blue respectively in the domain structure representations and sequence logos. In the sequence logos, larger residue characters indicate greater conservation of the residue across the FunFam.

2.3 Use of domain-based family resources for function annotation

The domain-centric approach can be exploited in functional annotation of the whole protein by identifying domains within a sequence, associating functions to these domains from the resource (eg Pfam, CATH) and integrating these functions in order to describe the function of the whole protein. Manually-curated GO associations for protein domain families are available for ProDom (ProDom2GO), Pfam (Pfam2GO) and InterPro (InterPro2GO) [51]. Various automated methods have been developed in recent years to exploit the functional signal encoded in domains to annotate uncharacterised proteins.

Schug and co-workers [52] developed a rule-based association of GO terms to ProDom [53] and CDD [40] domains for which thresholds were also determined. Query sequences were annotated by performing a BLAST search against ProDom or CDD followed by annotation transfer from matched domains that met the thresholds of domain-function associations. The GOtrees method [54] used decision trees to predict GO terms for query sequences based on domain composition in proteins (from Pfam) and other sequence features. Forslund and Sonnhammer [55] extended the Pfam2GO approach and developed two protocols: a rule-based (MultiPfam2GO) model that assigns a GO term to a domain if all proteins containing the domain are annotated with that GO term and a naïve Bayesian model, which associates GO terms to domains probabilistically. The SCOP2GO [56] method associates MFO terms to SCOP structural domains and annotates query sequences by scanning them against PSSM libraries that are built for SCOP domains having same fold and function (i.e. same GO terms). dcGO [57] or 'domain-centric GO' predictor infers GO terms for individual SCOP domains or supradomains (two or more domains which are known to function together) based on whole protein annotations from UniProtKB-GOA and domain architecture information extracted from SUPERFAMILY.

DFX [58] classifies the protein domain superfamilies in the CATH-Gene3D resource into domain functional families or FunFams using GO-based cluster evaluation of the hierarchical clustering algorithm, GeMMA (described earlier in section 2.2). Each FunFam is associated with GO terms probabilistically based on GO annotations of parent proteins of its domain sequences, which are then used to annotate query sequences based on their CATH domain composition. FunFHMMer [47] is an improved method for functional classification of CATH-Gene3D superfamilies which evaluates functional coherence of clusters using the evolutionary signals in cluster alignments and outperforms DFX and other domain-based classification protocols in predicting protein function. It can also be used to predict functionally important sites in query sequences as known functional sites have been found to be highly conserved in the FunFams generated by FunFHMMer (see Figure 2b).

Domain-based Prediction Method	Underlying Protein Domain Resource	Refs.

GO predictions from ProDom and CDD	ProDom and CDD	[52]
GOtrees	Pfam	[54]
MultiPfam2GO and probabilistic Naïve Bayesian model	Pfam	[55]
SCOP2GO	SCOP	[59]
dcGO	SCOP and SUPERFAMILY	[57]
DFX	CATH and Gene3D	[58]
FunFHMMer	CATH and Gene3D	[60,49]

Table 2: Protein function annotation methods which are based on protein domain families.

3. Function Annotation using FunFHMMer exploiting the CATH-Gene3D resource

CATH v4.0-Gene3D v12 identifies 110,439 FunFams for 2735 superfamilies. For the most populated FunFams, accounting for ~75% of CATH-Gene3D sequences, functionally important residues can also be predicted. All FunFam annotation data are made available through the CATH webpages (http://www.cathdb.info) (Figure 3). For each FunFam, the domain sequences are aligned using MAFFT [44], a profile hidden Markov model (HMM) is built using HMMER3 [61] and a model-specific threshold is determined. Each FunFam is then associated with a set of GO terms associated with the parent proteins of its annotated sequences.

Query sequences are scanned against the HMM models of the CATH FunFams and resolved into a single set of CATH domain architecture using DomainFinder3 [62]. Regions of the query sequences are assigned to a FunFam if they achieve the model-specific threshold and the GO terms associated with the FunFam are inherited by the query sequence along with a confidence score calculated by the frequency of each GO term among the annotated sequences of the particular FunFam. Finally, a non-redundant set of GO terms from all of the domain regions, each GO term retaining its highest confidence score, make up the GO annotations for the query sequence (Figure 4a).

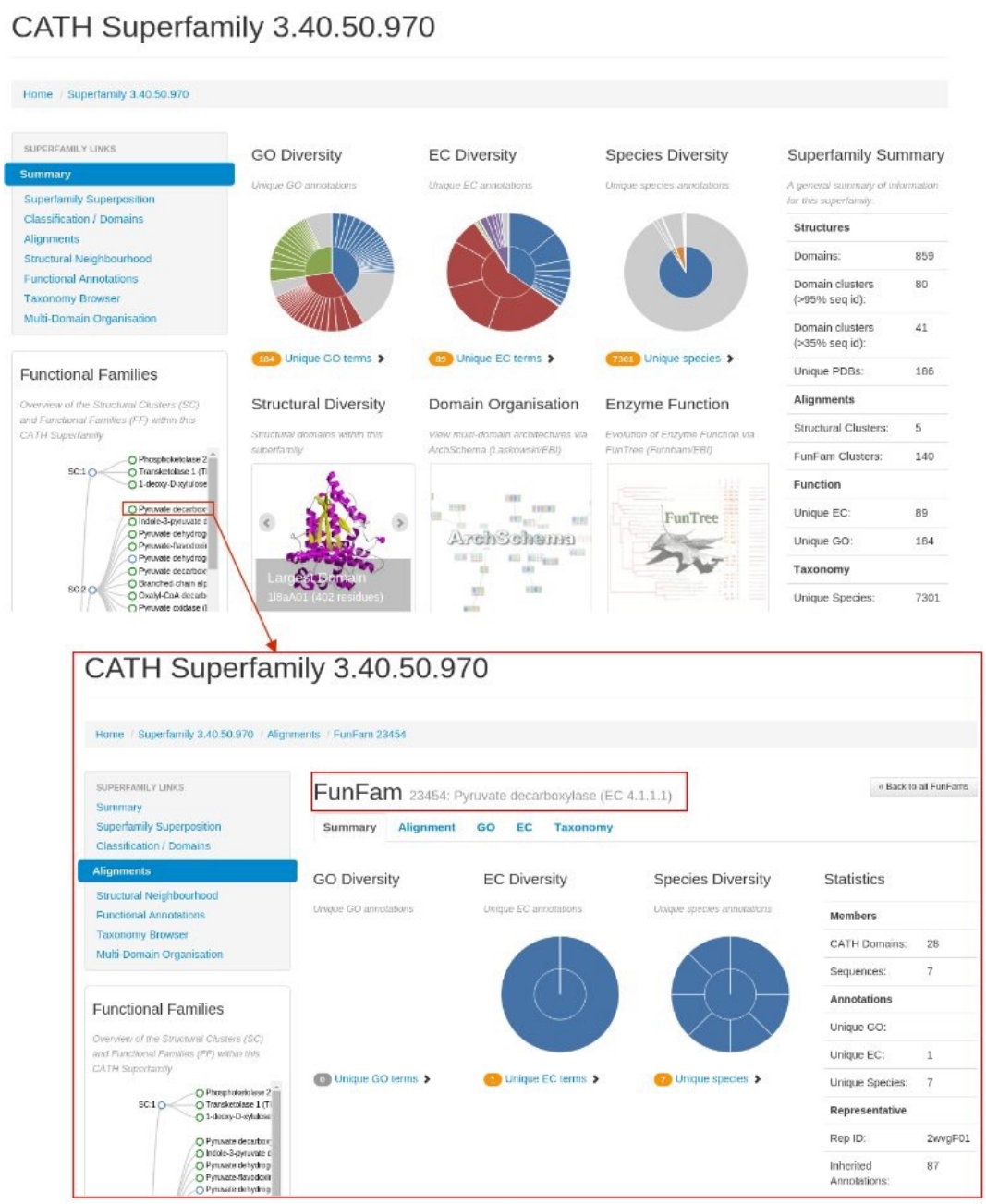

Figure 3: CATH webpages showing information on the Thiamine diphosphate (TPP)-dependant enzyme superfamily (CATH 3.40.50.970) and the Pyruvate decaboxylase FunFam within the TPP superfamily. The above webpages can be accessed from http://www.cathdb.info/superfamily/3.40.50.970.

3.1 Function annotation of uncharacterised sequences by CATH FunFams

The predictive power of the CATH FunFams have recently been evaluated using a rollback UniprotKB test set of 95 well-annotated proteins which had < 50% sequence identity to any experimentally annotated protein having GO molecular function ontology (MFO) terms [60]. FunFHMMer was found to perform much better than BLAST in this test set, where function annotation transfer from close homologs is limited (see Figure 4b for Precision-Recall graph as in CAFA [63]). Furthermore, the functional purity of the FunFams generated by FunFHMMer was also validated by CAFA 2, 2013-2014, a major bioinformatics initiative conducted by the Automated Function Prediction Special Interest Group (AFP-SIG), which aims to provide large-scale assessment of computational function prediction algorithms using a time challenge. In CAFA 2, a set of ~100,000 proteins lacking experimental annotations were provided to the automated function prediction community for submitting their predictions. After the submission deadline, the experimental annotations were allowed to accumulate over a period of 6 months and the prediction methods were evaluated on experimental annotations that had accumulated over the 6 month period. The preliminary results of CAFA 2 showed that FunFHMMer performed competitively, coming in the top 10 function prediction methods out of 110 methods in predicting Gene Ontology terms. CAFA 2 results can be accessed from: https://github.com/idoerg/CAFA2-results.

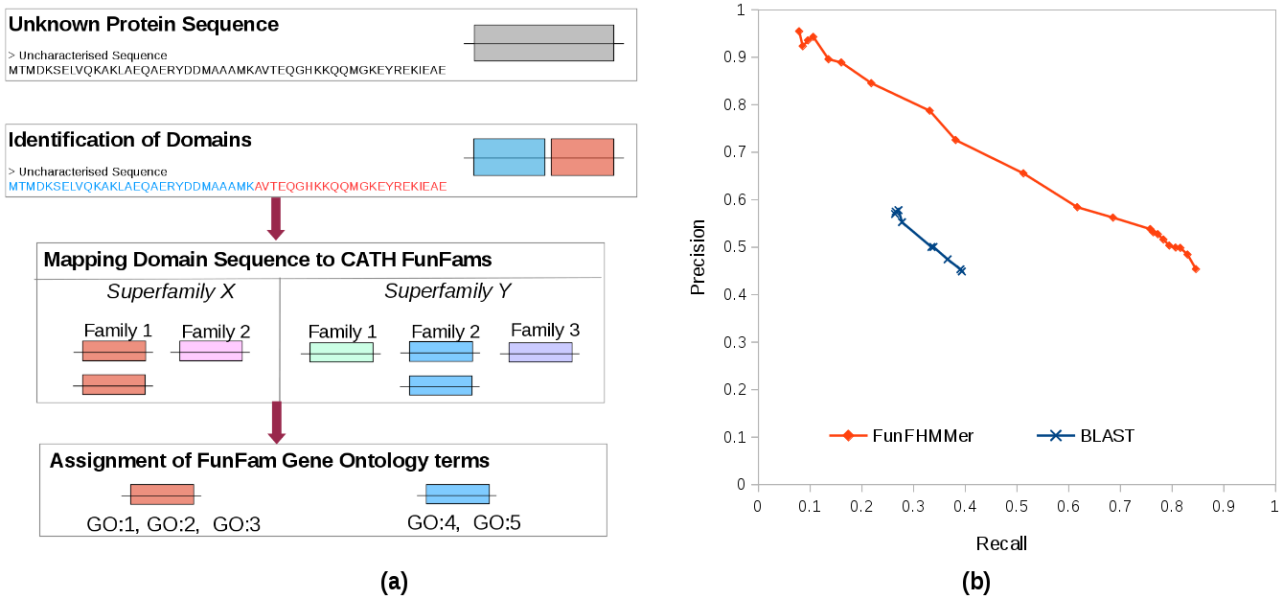

Figure 4: **(a)** Protocol for function prediction using the CATH FunFams. The multi-domain architecture (MDA) of the query protein sequence (shown as a grey box) is first identified. Two domains are identified in the query sequence (shown as blue and red boxes), which are then mapped to their closest CATH FunFam match. The GO annotations of the closest FunFam are then transferred to each of the domain regions, which together make up the GO annotations for the query sequence. **(b)** Precision-Recall graph showing the performance of FunFHMMer (in red) compared to BLAST (in blue)

3.2 Function Annotation of Moonlighting Proteins

A major challenge faced by computational approaches for protein function prediction protocols is the functional diversity of moonlighting proteins. Proteins which are capable of carrying out at least two diverse functions have been described as 'moonlighting'. For example, many glycolytic enzymes have been found to have a wide range of additional functions - from transcriptional repressors and chaperones to having virulence roles in many pathogens [64]. So far, the alternative function(s) of moonlighting proteins have been mostly discovered by serendipity and very little is known about the molecular mechanisms of proteins. They are known to switch their functions as a consequence of different cellular localization, cell type, oligomeric state, or cellular concentration of molecules. For example, Phosphoglucose isomerase functions as a glycolytic enzyme in the cytoplasm but as a nerve growth factor and cytokine outside the cell [10].

A number of existing computational tools have been analysed to determine whether current approaches for protein function prediction can disclose moonlighting functions of proteins [65,66]. Out of these, remote homology search by PSI-BLAST and profile-based search with Pfam were shown to have good performance for identifying moonlighting proteins [65]. PSI-BLAST results combined with information from protein-protein interaction (PPIs) databases were shown to give the best performance [66]. Recently, two methods by Kihara and co-workers, PFP [67] and ESG [68], have been shown to outperform PSI-BLAST [69], and are available as webservers. PFP (Protein Function Prediction) method uses a wide range of PSI-BLAST hits to query sequences to predict GO terms with several confidence measures utilizing data mining techniques. ESG (Extended Similarity Group) method performs iterative PSI-BLAST searches and predicts the function of a query sequence by combining information from even remote homologues to provide function annotations for a query protein with high reliability.

We investigated the performance of FunFHMMer in suggesting multi-functionality of proteins. We used a dataset of 144 proteins from the database of moonlighting proteins, MultitaskProtDB [70] to see whether the function annotations from CATH functional families can be used to suggest the multi-functionality of these proteins. All analyses were performed on the SwissProt database and GOA database dated November 2013 (considering only non-IEA GO terms). The performance of FunFHMMer on the moonlighting protein dataset was benchmarked against PSI-BLAST, BLAST and Pfam families, since PSI-BLAST and Pfam were shown in previous studies to perform well in predicting the moonlighting functions of proteins. PSI-BLAST was performed with the default setting of three iterations. Then all hits with an E-value score < 0.01 that have annotations, were used for transferring annotations to the query sequence. The GO term predictions were labelled according to the annotation frequency of a particular GO term amongst the PSI-BLAST hits and propagated up the tree. For the Pfam and FunFHMMer predictions, the moonlighting predictions were removed from the seed sequences of the respective Pfam families or CATH FunFams and their corresponding HMMs were then generated. The moonlighting proteins were then scanned against the HMMs and the GO terms of their FunFam top hits (E-value <0.01) were transferred to the query in a probabilistic manner calculated as the annotation frequency in a

matched family and propagated up the GO tree.

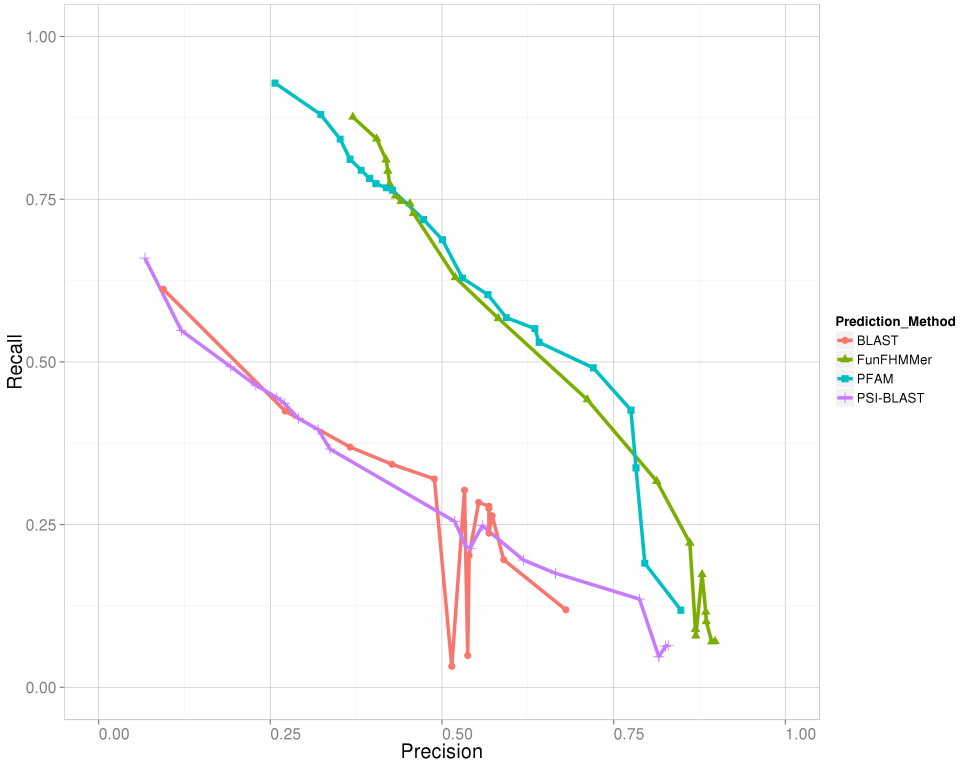

Figure 5: Comparison of the performance of FunFHMMer with PSI-BLAST, BLAST and Pfam-A in prediction of moonlighting proteins.

Performance of function predictions made by FunFHMMer compared with PSI-BLAST (number of iterations =3), BLAST and Pfam is illustrated in Figure 5 for Molecular Function Ontology (MFO) using a Precision-Recall curve as in CAFA [63]. The figure clearly indicates that both FunFHMMer and Pfam perform competitively and better than both BLAST and PSI-BLAST in predicting GO terms for the 144 moonlighting proteins in the dataset. Previous studies [65,66] have reported that methods aiming to detect diverse sequences (i.e. PSI-BLAST, PFP,ESG, or scans of Pfam families) can help in capturing the functional diversity of moonlighting proteins and aid in predicting secondary or alternative functions of these proteins, as these alternative functions are sometimes present in remote homologues. However, the FunFHMMer protocol is designed to predict functions based on functionally coherent FunFams, which are expected to distinguish between relatives which have any alternative functions when these are associated with different sequence motifs.

For example, the Chaperonin 60 apical domain (CATH 3.50.7.10) sequences for *Homo sapiens* and *Enterobacter aerogenes* which have two different moonlighting functions [71] are split into two different FunFams (3979 and 3904 respectively) in CATH v4.0 FunFams for the apical

domain superfamily. Moreover, an analysis of the conserved residues of the FunFams showed that FunFHMMer had identified the moonlighting motif which was reported in the literature (see Figure 6). As a result, we propose that there can be two approaches to identify moonlighting or alternative functions of a protein - (i) Inference from known functions of remote homologs, which suffers from the disadvantage that it would be very difficult for a biologist to identify a correct alternative function out of the numerous predicted ones. (ii) Using a finer classification of close homologs (e.g. CATH FunFams) to identify moonlighting motifs, which can aid in identifying moonlighting function of proteins. This approach is not as comprehensive as the former approach but would be easier for biologists to interpret the results.

Figure 6: The known moonlighting motif (in green) in Human HSP60 sequence is highly conserved

in its best match family in CATH-Gene3D (FunFam 3904) in the Chaperonin 60 apical domain superfamily but it is absent in a closely related family (FunFam 3979) containing bacterial sequences which have a different moonlighting activity.

3.3 FunFHMMer web server

The FunFHMMer web server is available at http://www.cathdb.info/search/by_funfhmmer [60]. The FunFHMMer web server can be queried using a protein sequence in the FASTA format or by entering UniProt/GenBank sequence identifiers as input in the text area on the webpage. A fully documented application programming interface (API) is also provided in the webserver to allow interfacing the FunFHMMer search from within any software application. The output of the web server provides the MDA of the query sequence along with CATH domain superfamily and FunFam assignments for each domain identified within the query sequence. The EC and GO annotations for each of the predicted FunFams are displayed in tables (see Figure 7).

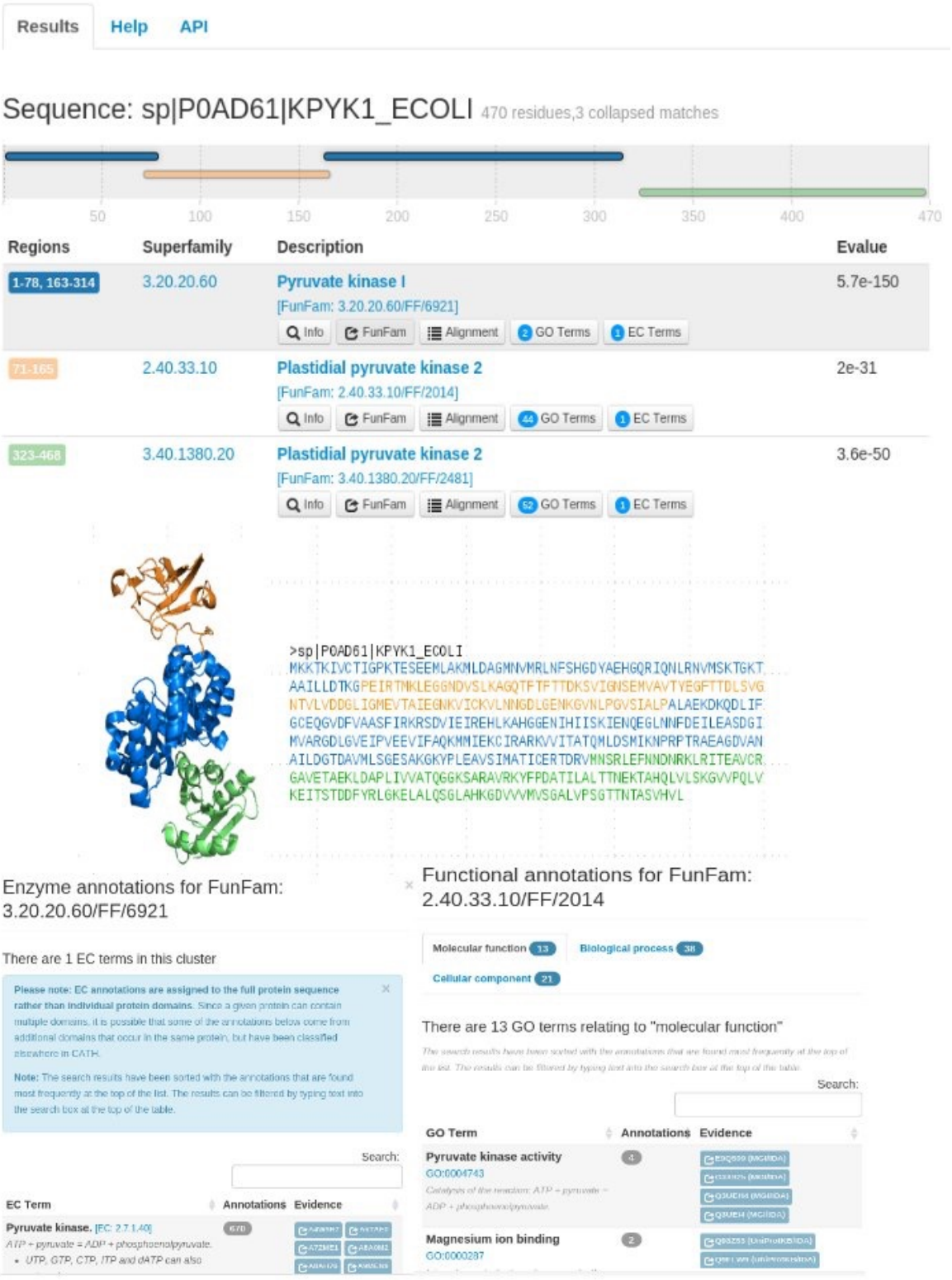

Figure 7: Functional annotations for query sequences provided by the FunFHMMer web server.

3.4 Visualization of Functional Family relationships

For each functionally diverse CATH superfamily (ie having two or more FunFams) the CATH website displays a cytoscape visualisation of the superfamily functional network (Figure 8), where functional families are represented by nodes and the edge distances correspond to the sequence similarity between the functional families. These can be very useful for understanding how function has been modulated by sequence or structure changes between functional families (FunFams) in a superfamily (see Figure 3). These networks help in providing a comprehensive summary of sequence, structure and function relationships in a functionally diverse superfamily which can aid in the identification of potentially novel targets for experimental characterization or structure determination eg by the structural genomics initiatives.

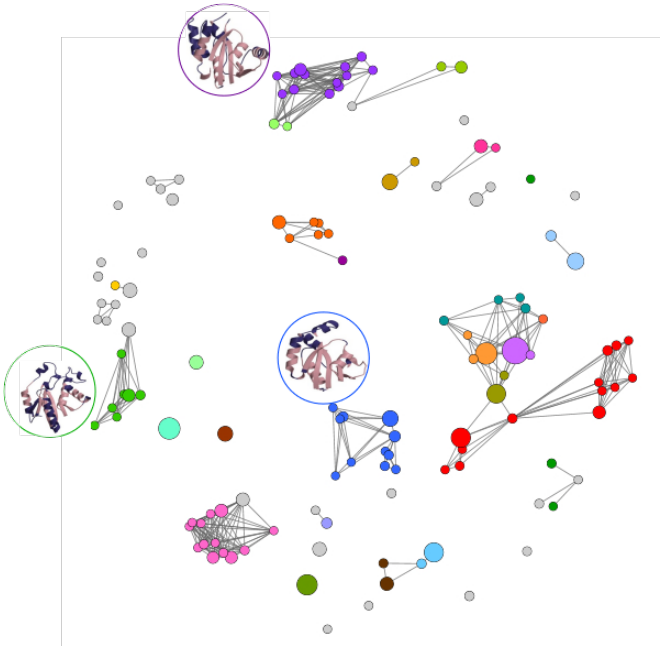

Figure 8: Visualization of sequence-structure-function relationships in a CATH superfamily (3.40.50.620) using Cytoscape v3.1 [72]. Each node corresponds to a FunFam which are coloured according to their enzyme classifications in the EC database. FunFams are linked if the similarity of their HMMs calculated by Profile Comparer (PRC) [73] are within a threshold PRC score of 50. For those FunFams having a structural representative, this is shown as an image in the figure.

4. Discussion / Challenges

Protein function is context-based and can be studied from different aspects: ranging from biochemical activity to the role of the protein in pathways, cells, tissues and organisms. A function annotation method using family resources is often limited by the scope of the family resources and their ability to provide functional information only for certain aspects. Moreover, bias in protein function annotations [74] or mis-annotations affects our understanding of protein function space [11]. As a result, sometimes correct and highly specific predictions may be misinterpreted as incorrect or erroneous if they have only been experimentally annotated in a generic manner. For example, annotated only as 'protein binding' rather than a more specific annotation term like 'tumor necrosis factor binding'.

Whilst the recent independent assessment (CAFA [63]) of methods for function prediction have been extremely valuable for determining which approaches work well, they have also shown how much more work needs to be done in providing reliable, accurate predictions [72]. Interestingly, in both CAFA1 and CAFA2 assessments, methods relying purely on whole protein or domain homology were amongst the top performing methods, sometimes outperforming machine learning methods that combined multiple additional information e.g. gene expression, cellular localisation. This suggests that there is considerable signal in the sequence reflecting the protein's molecular function and the context in which it operates. In this review, we have outlined several approaches for exploiting whole protein and domain homology to infer protein functions and shown the benefits of sub-classifying domain families into functional families to increase the accuracy of function prediction.

References

[1] T.B.K. Reddy, A.D. Thomas, D. Stamatis, J. Bertsch, M. Isbandi, J. Jansson, et al., The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification., Nucleic Acids Res. (2014) gku950–. doi:10.1093/nar/gku950.

[2] The UniProt Consortium, UniProt: a hub for protein information, Nucleic Acids Res. (2014) gku989–. doi:10.1093/nar/gku989.

[3] W.A. Baumgartner, K.B. Cohen, L.M. Fox, G. Acquaah-Mensah, L. Hunter, Manual curation is not sufficient for annotation of genomic databases, Bioinformatics. 23 (2007) i41–i48. doi:10.1093/bioinformatics/btm229.

[4] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[5] O.C. Redfern, A. Harrison, T. Dallman, F.M.G. Pearl, C.A. Orengo, CATHEDRAL: A Fast and Effective Algorithm to Predict Folds and Domain Boundaries from Multidomain Protein Structures, PLoS Comput Biol. 3 (2007) e232+. doi:10.1371/journal.pcbi.0030232.

[6] S. Addou, R. Rentzsch, D. Lee, C. a Orengo, Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer., J. Mol. Biol. 387 (2009) 416–30. doi:10.1016/j.jmb.2008.12.045.

[7] S. Erdin, A.M. Lisewski, O. Lichtarge, Protein function prediction: towards integration of similarity metrics, Curr. Opin. Struct. Biol. 21 (2011) 180–188.

[8] A.C. Martin, C.A. Orengo, E.G. Hutchinson, S. Jones, M. Karmirantzou, R.A. Laskowski, et al., Protein folds and functions, Structure. 6 (1998).

[9] M. Bashton, C. Chothia, The generation of new protein functions by the combination of domains, Structure. 15 (2007) 85–99.

[10] C. Jeffery, M. Mani, V. Amblee, C. Chen, Moonlighting Proteins, Biophys. J. 106 (2014). doi:10.1016/j.bpj.2013.11.3640.

[11] A.M. Schnoes, S.D. Brown, I. Dodevski, P.C. Babbitt, Annotation error in public databases: misannotation of molecular function in enzyme superfamilies., PLoS Comput. Biol. 5 (2009) e1000605. doi:10.1371/journal.pcbi.1000605.

[12] J.A. Gerlt, P.C. Babbitt, M.P. Jacobson, S.C. Almo, Divergent Evolution in Enolase Superfamily: Strategies for Assigning Functions, J. Biol. Chem. 287 (2012) 29–34. doi:10.1074/jbc.r111.240945.

[13] B.H. Dessailly, O.C. Redfern, A.L. Cuff, C. a Orengo, Detailed analysis of function divergence in a large and diverse domain superfamily: toward a refined protocol of function classification., Structure. 18 (2010) 1522–1535. doi:10.1016/j.str.2010.08.017.

[14] M. Bashton, C. Chothia, The generation of new protein functions by the combination of domains., Structure. 15 (2007) 85–99. doi:10.1016/j.str.2006.11.009.

[15] S.S. Hannenhalli, R.B. Russell, Analysis and prediction of functional sub-types from protein sequence alignments, J. Mol. Biol. 303 (2000) 61–76.

[16] A.E. Todd, C.A. Orengo, J.M. Thornton, Evolution of function in protein superfamilies, from a structural perspective, J. Mol. Biol. 307 (2001) 1113–1143. doi:10.1006/jmbi.2001.4513.

[17] H. Mi, A. Muruganujan, P.D. Thomas, PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees, Nucleic Acids Res. 41 (2013) D377–D386.

[18] D.H. Haft, J.D. Selengut, O. White, The TIGRFAMs database of protein families, Nucleic Acids Res. 31 (2003) 371–373.

[19] I. Pedruzzi, C. Rivoire, A.H. Auchincloss, E. Coudert, G. Keller, E. De Castro, et al., HAMAP in 2013, new developments in the protein family classification and annotation system, Nucleic Acids Res. 41 (2013) D584–D589.

[20] N. Krishnamurthy, D. Brown, D. Kirshner, K. Sjölander, PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification, Genome Biol. 7 (2006) R83.

[21] K. Sjolander, Phylogenetic inference in protein superfamilies: analysis of SH2 domains, in: Proc. Int. Conf. Intell. Syst. Mol. Biol, 1998: pp. 165–174.

[22] D.P. Brown, N. Krishnamurthy, K. Sjo, Automated Protein Subfamily Identification and Classification, 3 (2007). doi:10.1371/journal.pcbi.Citation.

[23] N. Wicker, G.R. Perrin, J.C. Thierry, O. Poch, Secator: a program for inferring protein subfamilies from phylogenetic trees, Mol. Biol. Evol. 18 (2001) 1435–1441.

[24] N. Rappoport, N. Linial, M. Linial, ProtoNet: charting the expanding universe of protein sequences., Nat. Biotechnol. 31 (2013) 290–2. doi:10.1038/nbt.2553.

[25] N. Rappoport, A. Stern, N. Linial, M. Linial, Entropy-driven partitioning of the hierarchical protein space, Bioinformatics. 30 (2014) i624–i630. doi:10.1093/bioinformatics/btu478.

[26] R. Petryszak, E. Kretschmann, D. Wieser, R. Apweiler, The predictive power of the CluSTr database, Bioinformatics. 21 (2005) 3604–3609.

[27] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E. V Koonin, et al., The COG database: an updated version includes eukaryotes., BMC Bioinformatics. 4 (2003) 41. doi:10.1186/1471-2105-4-41.

[28] L. Li, C.J. Stoeckert, D.S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes, Genome Res. 13 (2003) 2178–2189.

[29] S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, et al., eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges., Nucleic Acids Res. 40 (2012) D284–9. doi:10.1093/nar/gkr1060.

[30] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, et al., Gene Ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29.

[31] D. Piovesan, P.L. Martelli, P. Fariselli, G. Profiti, A. Zauli, I. Rossi, et al., How to inherit statistically validated annotation within BAR+ protein clusters., BMC Bioinformatics. 14 Suppl 3 (2013) S4. doi:10.1186/1471-2105-14-S3-S4.

[32] R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, et al., Pfam: the protein families database, Nucleic Acids Res. 42 (2014) D222–D230.

[33] A. Andreeva, D. Howorth, J.-M. Chandonia, S.E. Brenner, T.J.P. Hubbard, C. Chothia, et al., Data growth and its impact on the SCOP database: new developments, Nucleic Acids Res. 36 (2007) D419–D425. doi:10.1093/nar/gkm993.

[34] M.E. Oates, J. Stahlhacke, D. V Vavoulis, B. Smithers, O.J.L. Rackham, A.J. Sardar, et al., The SUPERFAMILY 1.75 database in 2014: a doubling of data., Nucleic Acids Res. 43 (2015) D227–33. doi:10.1093/nar/gku1041.

[35] I. Sillitoe, T.E. Lewis, A. Cuff, S. Das, P. Ashford, N.L. Dawson, et al., CATH: comprehensive structural and functional annotations for genome sequences, Nucleic Acids Res. 43 (2015) D376–D381. doi:10.1093/nar/gku947.

[36] J.G. Lees, D. Lee, R.A. Studer, N.L. Dawson, I. Sillitoe, S. Das, et al., Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis, Nucleic Acids Res. 42 (2014) D240–D245.

[37] H. Cheng, R.D. Schaeffer, Y. Liao, L.N. Kinch, J. Pei, S. Shi, et al., ECOD: An Evolutionary Classification of Protein Domains, PLoS Comput. Biol. 10 (2014) e1003926.

[38] S. Abhiman, E.L.L. Sonnhammer, FunShift: a database of function shift analysis on protein subfamilies, Nucleic Acids Res. 33 (2005) D197–D200.

[39] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T.K. Attwood, A. Bateman, et al., InterPro in 2011: new developments in the family and domain prediction database., Nucleic Acids Res. 40 (2012) D306–D312. doi:10.1093/nar/gkr948.

[40] A. Marchler-Bauer, C. Zheng, F. Chitsaz, M.K. Derbyshire, L.Y. Geer, R.C. Geer, et al., CDD: conserved domains and protein three-dimensional structure, Nucleic Acids Res. 41 (2013) D348–D352. doi:10.1093/nar/gks1243.

[41] A. Cuff, O.C. Redfern, L. Greene, I. Sillitoe, T. Lewis, M. Dibley, et al., The CATH Hierarchy Revisited — Structural Divergence in Domain Superfamilies and the Continuity of Fold Space, Struct. Des. 17 (2009) 1051–1062. doi:10.1016/j.str.2009.06.015.

[42] D.A. Lee, R. Rentzsch, C. Orengo, GeMMA : functional subfamily classification within superfamilies of predicted protein structural domains, 38 (2010) 720–737. doi:10.1093/nar/gkp1049.

[43] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-

generation sequencing data, Bioinformatics. 28 (2012) 3150–3152.

[44] K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 30 (2002) 3059–3066.

[45] R. Sadreyev, N. Grishin, COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance, J. Mol. Biol. 326 (2003) 317–336.

[46] R. Rentzsch, C.A. Orengo, Protein function prediction–the power of multiplicity, Trends Biotechnol. 27 (2009).

[47] S. Das, D. Lee, I. Sillitoe, N. Dawson, J. Lees, C. Orengo, Functional classification of CATH superfamilies: a domain-based approach for protein function annotation, Bioinformatics. btv398 (2015).

[48] E. Akiva, S. Brown, D.E. Almonacid, A.E. Barber, A.F. Custer, M.A. Hicks, et al., The structure--function linkage database, Nucleic Acids Res. 42 (2013) D521–D530.

[49] S. Das, Functional Sub-classification of Domain Superfamilies (Upgrade Report), 1 (2014) 1–21.

[50] W.S.J. Valdar, Scoring residue conservation, Proteins Struct. Funct. Bioinforma. 48 (2002) 227–241.

[51] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, et al., The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology., Nucleic Acids Res. 32 (2004) D262–6. doi:10.1093/nar/gkh021.

[52] J. Schug, S. Diskin, J. Mazzarelli, B.P. Brunk, C.J. Stoeckert, Predicting gene ontology functions from ProDom and CDD protein domains., Genome Res. 12 (2002) 648–55. doi:10.1101/gr.222902.

[53] C. Bru, E. Courcelle, S. Carrère, Y. Beausse, S. Dalmar, D. Kahn, The ProDom database of protein domain families: more emphasis on 3D., Nucleic Acids Res. 33 (2005) D212–5. doi:10.1093/nar/gki034.

[54] B. Hayete, J.R. Bienkowska, Gotrees: predicting go associations from protein domain composition using decision trees., in: Pacific Symp. Biocomput., 2005: pp. 127–138.

[55] K. Forslund, E.L.L. Sonnhammer, Predicting protein function from domain content., Bioinformatics. 24 (2008) 1681–7. doi:10.1093/bioinformatics/btn312.

[56] D. Lopez, F. Pazos, Concomitant prediction of function and fold at the domain level with GO-based profiles., BMC Bioinformatics. 14 Suppl 3 (2013) S12. doi: 10.1186/1471-2105-14-S3-S12.

[57] H. Fang, J. Gough, A domain-centric solution to functional genomics via dcGO Predictor., BMC Bioinformatics. 14 (2013) S9. doi:10.1186/1471-2105-14-S3-S9.

[58] R. Rentzsch, C.A. Orengo, Protein function prediction using domain families, BMC

Bioinformatics. 14 (2013) S5.

[59] D. Lopez, F. Pazos, Concomitant prediction of function and fold at the domain level with GO-based profiles., BMC Bioinformatics. 14 Suppl 3 (2013) S12. doi: 10.1186/1471-2105-14-S3-S12.

[60] S. Das, I. Sillitoe, D. Lee, J.G. Lees, N.L. Dawson, J. Ward, et al., CATH FunFHMMer web server: protein functional annotations using functional family assignments, Nucleic Acids Res. 43 (2015) W148–153. doi:10.1093/nar/gkv488.

[61] S.R. Eddy, others, A new generation of homology search tools based on probabilistic inference, in: Genome Inf., 2009: pp. 205–211.

[62] C. Yeats, O.C. Redfern, C. Orengo, A fast and automated solution for accurately resolving protein domain architectures., Bioinformatics. 26 (2010) 745–751. doi: 10.1093/bioinformatics/btq034.

[63] P. Radivojac, W.T. Clark, T.R. Oron, A.M. Schnoes, T. Wittkop, A. Sokolov, et al., A large-scale evaluation of computational protein function prediction, Nat. Methods. 10 (2013) 221–227. doi:10.1038/nmeth.2340.

[64] B. Henderson, A. Martin, Bacterial virulence in the moonlight: multitasking bacterial moonlighting proteins are virulence determinants in infectious disease., Infect. Immun. 79 (2011) 3476–91. doi:10.1128/IAI.00179-11.

[65] A. Gómez, S. Hernández, I. Amela, J. Piñol, J. Cedano, E. Querol, Do protein-protein interaction databases identify moonlighting proteins?, Mol. Biosyst. 7 (2011) 2379–82. doi:10.1039/c1mb05180f.

[66] S. Hernández, L. Franco, A. Calvo, G. Ferragut, A. Hermoso, I. Amela, et al., Bioinformatics and Moonlighting Proteins, Front. Bioeng. Biotechnol. 3 (2015). doi: 10.3389/fbioe.2015.00090.

[67] T. Hawkins, M. Chitale, S. Luban, D. Kihara, PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data., Proteins. 74 (2009) 566–82. doi:10.1002/prot.22172.

[68] M. Chitale, T. Hawkins, C. Park, D. Kihara, ESG: extended similarity group method for automated protein function prediction., Bioinformatics. 25 (2009) 1739–45. doi: 10.1093/bioinformatics/btp309.

[69] S. Altschul, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402. doi:10.1093/nar/ 25.17.3389.

[70] S. Hernández, G. Ferragut, I. Amela, J. Perez-Pons, J. Piñol, A. Mozo-Villarias, et al., MultitaskProtDB: a database of multitasking proteins., Nucleic Acids Res. 42 (2014) D517–20. doi:10.1093/nar/gkt1153.

[71] B. Henderson, M.A. Fares, P.A. Lund, Chaperonin 60: a paradoxical, evolutionarily conserved protein family with multiple moonlighting functions., Biol. Rev. Camb.

Philos. Soc. 88 (2013) 955–87. doi:10.1111/brv.12037.

[72] M. Kohl, S. Wiese, B. Warscheid, Cytoscape: software for visualization and analysis of biological networks, in: Data Min. Proteomics, Springer, 2011: pp. 291–303.

[73] M. Madera, Profile Comparer: a program for scoring and aligning profile hidden Markov models., Bioinformatics. 24 (2008) 2630–1. doi:10.1093/bioinformatics/btn504.

[74] A.M. Schnoes, D.C. Ream, A.W. Thorman, P.C. Babbitt, I. Friedberg, Biases in the experimental annotations of protein function and their effect on our understanding of protein function space., PLoS Comput. Biol. 9 (2013) e1003063. doi:10.1371/journal.pcbi.1003063.