Chapter 5

Protein Function Prediction

Leonardo Magalhães Cruz, Sheyla Trefflich, Vinícius Almir Weiss, and Mauro Antônio Alves Castro

Abstract

Protein function is a concept that can have different interpretations in different biological contexts, and the number and diversity of novel proteins identified by large-scale "omics" technologies poses increasingly new challenges. In this review we explore current strategies used to predict protein function focused on high-throughput sequence analysis, as for example, inference based on sequence similarity, sequence composition, structure, and protein–protein interaction. Various prediction strategies are discussed together with illustrative workflows highlighting the use of some benchmark tools and knowledge bases in the field.

Key words Protein function, Homology, Ontology, Biological databases, Database sequence similarity search, Protein families, Protein domains, Phylogeny, Bioinformatics

1 Introduction

With the advent of structural, functional, and comparative genomics, numerous sequences of predicted proteins have been produced in a velocity that cannot be followed by its experimental studies, and the only feasible way to annotate tentative functions to these proteins is by means of automatic sequence analysis [1]. Beyond sequence, structural genomic projects have also allowed the determination of protein structure in a high-throughput fashion [2]. On the other hand, although these methodologies contribute to our knowledge, over one-third of structures are of proteins of unknown function and their worth can only be significantly enhanced by knowing the biological roles that they play [2], but experimental characterization of function cannot scale up to accommodate the vast amount of sequence [3] and structural data already available and the growing gap between sequences and experimentally annotated proteins can only be accomplished by combining experimental and computational methods for functional annotation [4]. Further, experimental efforts have been

Michael Kaufmann et al. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1654,
DOI 10.1007/978-1-4939-7231-9_5,

done to determine protein function and provide a more detailed understanding mainly of model organisms, expecting that accurate annotation may be transferred to other species by computational methods [4]. Numerous approaches have been used to automatically predict protein function so far, from different data types, such as sequence information, protein structure, phylogenetics and evolutionary relationships, interaction and association data, and a combination of these [5].

The accurate annotation of protein function is a key to understanding life at the molecular level and has great biomedical and pharmaceutical implications [3, 4]. In the absence of experimental data, the function of a protein can be inferred on the basis of its sequence similarity, sequence composition, structure [3], gene expression, protein–protein interaction, phylogeny, genomic context, or other structural or functional information based on our knowledge about proteins with already known functions. Even in the presence of some experimental evidences, automatic analysis is important to integrate data and evidences for function, because experimental characterization of a protein such as structural data, analysis of gene expression, and delineation of a protein interaction network rarely gives direct clues to gene function [6]. The computational annotation of protein function has therefore emerged as a problem at the forefront of computational and molecular biology [3]. However, prediction of function from sequence is a considerably more complex enterprise than a simple sequence database search [7].

1.1 Homology

Similar genes often have conserved functions in different organisms. This happens because organisms share a common evolutionary history, preserving functions from a common ancestor and changing it along time of evolution. These shared functions or characteristics linked by a common ancestor is called homology and cannot be quantified. Functions or characteristics "are" or "are not" homologous. However, ancestral organisms or states are not present today and the way to infer homology is by means of quantification of similarity. For nucleotide and amino acid sequences, the way to measure similarity is by means of a sequence alignment. Different types of homologies may be distinguished and the main ones are (Fig. 1a): orthologs, arisen by an speciation evolutionary event, and paralogs, arisen by a gene duplication evolutionary event. Time of evolution may modify nucleotide or protein sequences and lengths, but it is also important to consider the evolution of proteins in another perspective. Many proteins are structured in a domain manner, meaning that these proteins are composed by a set of independent functional units and each of these domains may have a different evolutionary history.

The more the organisms evolve, the more the sequences diverge and the more difficult is it to establish similarity and infer homology from similarity and sequence alignments. This relationship can also

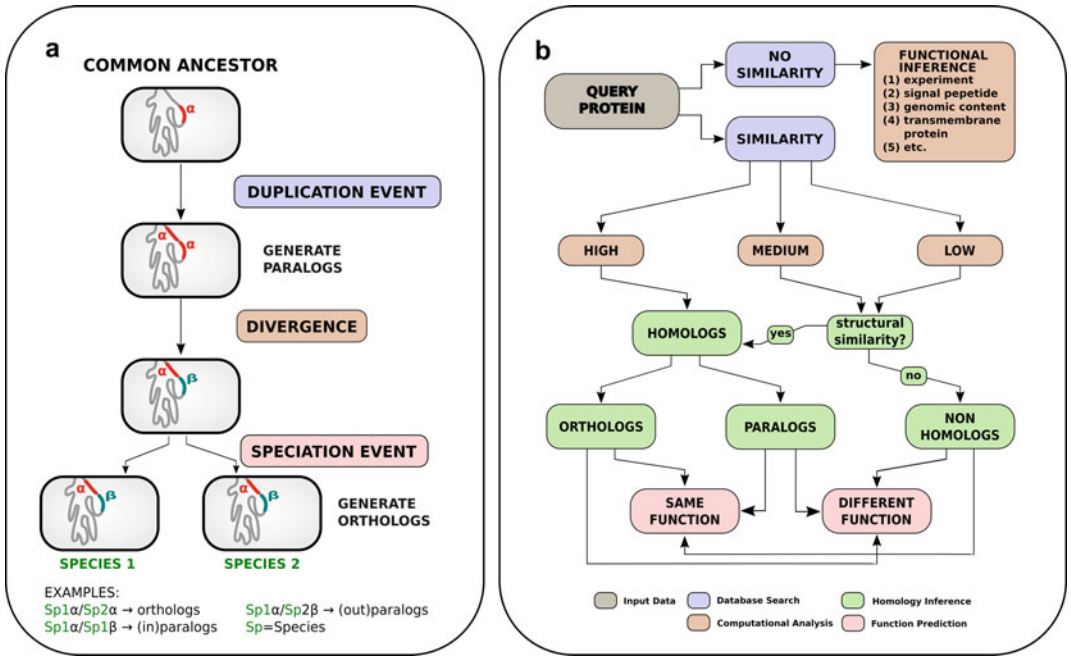

Fig. 1 Sequence similarity and homology in protein function prediction. Flowcharts summarizing (**a**) basic concepts on homology and sequence divergence and (**b**) possible strategies in protein annotation using sequence similarity

be used for protein function prediction, where the higher the sequence similarity, the better the chance that homologous proteins in fact share functional features [8]. For that purpose, the following rule may be useful [9]: (a) 90% of protein sequences sharing 30%, or more, identity are structurally similar, suggesting high probability of homology and also function; (b) only 10% of protein sequences sharing 25%, or less, sequence identity are structurally similar, suggesting a low probability to find homology and function.

Directly or indirectly, the prediction of a protein function in silico passes through the identification of homologous and the measurement of similarity that, at the end, will allow homology identification. On the contrary, displacement of non-homologous but functionally equivalent enzymes [7] is also observed.

1.2 Definition of Protein Function

Protein function is a concept that can have different interpretations in different biological contexts and/or level [8, 10–12], describing biochemical, cellular, and phenotypic aspects of the molecular events that involve the protein [3, 4]. The protein function can be divided into three major categories: (a) molecular function, e.g., the specific reaction catalyzed by an enzyme; (b) biological process, e.g., the metabolic pathway the enzyme is involved in; and (c) system or physiological level, e.g., if the enzyme is involved in respiration, photosynthesis, cell signaling, etc. One could also

consider a fourth level of cellular component, specifying the compartment of the cell the protein plays its role, e.g., cell membrane, any organelles [8, 11–13]. Protein function may also vary in space and time [11], as we will see, for example, in the case of moonlighting proteins. Computational methods exist to predict all of these aspects of function [13]. Furthermore, most biological processes are carried out by groups of interacting proteins and these interactions can be predicted in silico [13]. These many levels of protein function, from a very specific biochemical activity to a biological processes and pathways context, and from the cell to the organism level [2] generate practical consequences with protein annotation including vague terms to describe its function, such as "like protein," "containing domain protein," and "signaling protein" [2].

When attempting to identify the molecular function of a protein, it is important to bear in mind the simple rule: sequence → structure → function, that is, sequence determines the structure and structure determines the molecular function.

When describing function, attention must be paid to two kinds of proteins: those containing multiple domains and the so calling moonlighting proteins. The former are proteins composed of many domains, each domain contributing with a different specialized function to compose a unique biological function of the protein. Variation in the domain composition may occur, given different functions to similar proteins within the same family. The last are proteins that perform more than one function (multitask protein). For a moonlighting protein, usually independent unrelated functions are observed [14], not including function variation that results from gene fusions, homologous but nonidentical proteins, proteins resulting from alternative splicing, variation in posttranslational modifications and proteins operating in different locations or are able to utilize different substrates but have a single function [15].

It is now recognized that multifunctional proteins are common [4]. At least 34% of functionally characterized proteins (by experimental studies) are already assigned more than one distinct molecular function term and that at least 56% of proteins participate in more than one distinct biological process [4].

Different function of moonlighting proteins occur due to [15]: (a) cellular localization (within the cell or if inside/outside the cell); (b) the cell types expressing the protein; (c) the substrate, product, or a cofactor bound to the protein or different binding sites for different ligands; (d) the number of subunits joined and variation in the complexes to form the quaternary structure of a protein. These mechanisms that a protein can moonlight demonstrate the function may shift at different levels (i.e., molecular function, cellular process, or localization). The MoonProt database actually lists approximately 300 experimentally identified moonlighting proteins (www.moonlightingproteins.org).

If the moonlighting functions of a protein may also be assigned to an unknown protein by means of homology-based transfer is a matter of discussion. Identification of additional function of moonlight proteins is relatively recent and difficult by experimentation and its identification by in silico analysis is an even greater challenge [14]. Few methods are actually available to predict moonlighting proteins. Khan et al. [14] searched GO for known moonlighting proteins and observed that clusters of these proteins reflect their functions. Further analysis of protein–protein interaction, gene expression, phylogenetic profile, and genetic interaction network revealed that moonlighting proteins physically interact with a higher number of distinct functional classes of proteins than non-moonlighting proteins and that moonlighting proteins tend to interact with other moonlighting proteins. It has also been suggested that moonlighting proteins are under positive selection [14, 15]. These observations open the door for in silico prediction of moonlighting functions.

1.3 Proteins of Unknown Function

A large portion of known proteins are poorly characterized experimentally, with very little knowledge about their function [8]. The vast majority of proteins with function experimentally verified is observed in model organisms [4], but even for those organisms, a significant part of all proteins coded in their genomes are to be characterized. In *Escherichia coli* K-12, about one-third (1408) of the 4225 predicted proteins remain functionally unannotated (orphans) and only half of the predicted proteins have indicative of function based on experimental evidence and the same proportion seems to apply to *Saccharomyces cerevisiae* [6, 16]. Further, the remaining genes between experimentally annotated and unannotated in *E. coli* have either only generic functional attributes [16].

In Swiss-Prot v15.15, a curated database, approximately 90% of annotated proteins in Molecular Function and Biological Process ontologies belong to nine model organisms only (*H. sapiens*, *S. cerevisiae*, *M. musculus*, *R. norvegicus*, *A. thaliana*, *D. melanogaster*, *S. pombe*, *E. coli* K-12, and *C. elegans*) [4]. However, nearly 60% of the proteins from these model organisms still do not have any experimentally determined Molecular Function or Biological Process terms [4].

In CharProtDB (www.jcvi.org/charprotdb) [17], a database of experimentally characterized proteins, updated dataset till 2011 indicate that the main organisms with experimentally characterized proteins are as follow: *Escherichia coli* with 2631 proteins (~60% of all proteins), *Schizosaccharomyces pombe* with 1817 proteins (~35%), *Candida albicans* with 1308 proteins (~9%), and *Bacillus subtillis* with 1250 proteins (~30%). A total of 1252 species of all domain of life are included in the database and 96% of them have less than 100 experimentally characterized proteins.

Although these information about experimentally characterized proteins is difficult to obtain and is presented from different source and time, taken together, they give us an overview of our current knowledge about the function of proteins in different organisms and our need for tools that allow of automatic and reliable prediction of protein function.

In Pfam (pfam.xfam.org) [18] release 26.0, a database dedicated to protein families and its domains, more than 20% of all proteins are annotated as containing DUFs (Domains of Unknown Function) [19]. A total of 355 essential proteins in 16 model bacterial species contain 238 DUFs, most of which represent single-domain proteins, clearly establishing the biological essentiality of DUFs [19]. About 9% of DUFs spanned all domains of life, nearly half (43%) had been detected only in bacteria, 19% were only found in eukaryotes, and 3% are restricted to Archaea [20].

For the updated version of COG (Clusters of Orthologous Groups; www.ncbi.nlm.nih.gov/COG) [1], a database of putative orthologous proteins shared from completely sequenced genomes of bacteria and archaea, among a total of 4631 COGs distributed in 26 functional categories, R "General function prediction only" (507 COGs) and S "Function unknown" (959 COGs) are the most abundant categories, both counting for 31.6% of all COGs. Further, all COGs include about 60% and 86% of bacterial and archaeal proteomes, respectively [1], with remaining proteins not even being assigned to any existing COG. The fraction of the total proteome with specific functional annotation (excluding R and S categories) varies from a minimum of about 51–53% to a maximum of 72–76% at the phyla level [1].

The large number of functionally unannotated genes is observed because experimental characterization is time consuming, so these genes have never been studied experimentally or experimental studies brought contradictory results that could not be easily reconciled [6].

2 Strategies for Protein Function Prediction

Normally, the prediction of a protein function starts by trying to define its molecular function, using a homology-based transfer strategy, e.g., a similarity search against a database of known proteins or a search against a protein family and domain database. In a next step, one tries to extend the molecular function to a system function, that is, define the role played by a protein in a biological process.

Computational biology offers tools that can provide insight into the function of proteins based on their sequence, their structure, their evolutionary history, and their association with other proteins [8]. There are also methods that directly analyze the

sequence or structure in order to predict the function or methods that rely on sources of information that are beyond the protein itself, such as genomic context, protein–protein interaction networks, or membership in biochemical pathways [8].

Prediction of protein function, unlike establishing homology, is not a "yes" or "no" decision (i.e., an unknown protein will or will not have exactly the same function than a homologous counterpart). Function may be shared at different levels. The obvious example is two proteins that participate in the same cellular process but have different enzymatic activities (i.e., share the same cellular process function but have different molecular functions). Further, if two proteins are homologous, it means that they share a common evolutionary origin, but it does not guarantee that these two proteins will have the same function [8]. On the other hand, concerning about different kinds of homology, in general, functions from ancestral origin tend to be conserved more in orthologs than in paralogs [8, 21], but frequently distinguishing between them is not a straightforward task and even orthologs may diverge functionally [8, 21]. In the opposite way, proteins with same function may arise not by means of homology, but by convergent evolution, when by means of adaptive change, some molecular "functionality" arises independently in proteins not sharing an ancestral sequence [22, 23]. All these possibilities are presented in Fig. 1b, showing how homology, similarity, and function correlate.

Function predicted automatically and on a large scale includes additional problems concerning the need to standardize and quantitatively assess the similarity of functions between proteins [8]. A large number of methods have been proposed to predict protein function using information from amino acid sequence and predicted physicochemical properties, phylogenetic profiles and genomic context, protein–protein interaction networks, protein structure data, microarrays and clustering patterns of coregulated genes, predicted ligands, or a combination of data types [3, 4, 24].

The primary databases of biological sequences and structures are the main sources of information for any methods attempting to predict protein function. These databases can be directly searched to looking for similar sequences or structures and infer homology to transfer functional annotation or can be used to build secondary databases of clusters of protein sequences (e.g., COG, UniProtKB/UniRef, NCBI Protein Clusters, Panther), family and domains (e.g., Pfam, PROSITE, SMART, PRINTS, CDD), protein domain classification from structures and sequences (e.g., CATH, Gene3D), or retrieve well-known and annotated sequences/structures experimentally characterized to build probabilistic models or models based on machine learning that may be applied to scan unknown proteins to give insight in its function (e.g., TMHMM, LocTree3, BaCelLo, TargetP, PSORT, Protein prowler, LipoP, TatP). In this sense, all knowledge applied to automatically predict

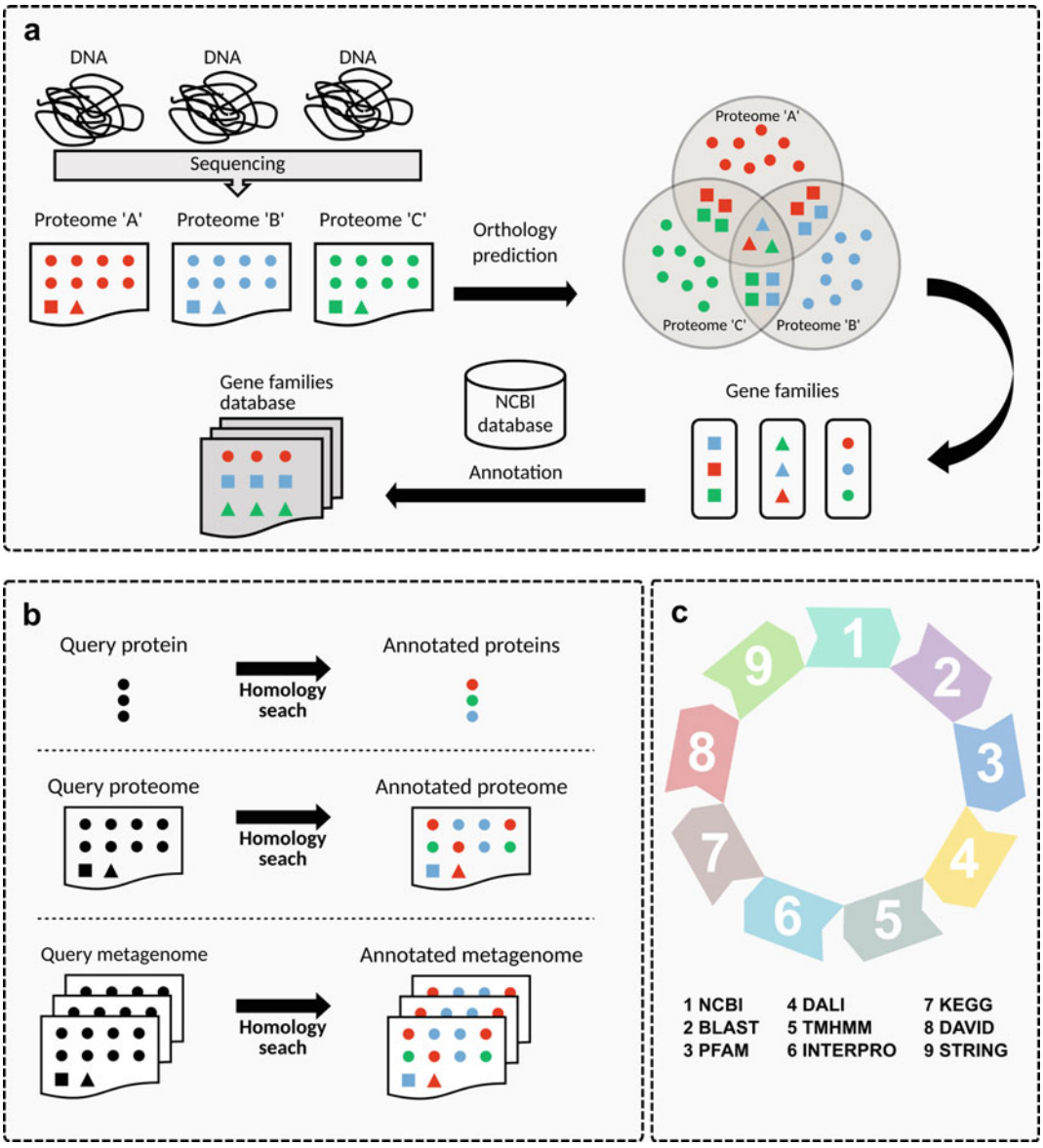

Fig. 2 Protein annotation strategies using knowledge bases. Flowcharts exemplifying (**a**) knowledge base construction and (**b**) the annotation process of a protein sequence, a proteome and a metagenome using homology-searching strategies. (**c**) The combination of different resources can be used for knowledge discovery in databases in order to help the annotation process (*see* Fig. 3 for additional details)

the function of a protein from its sequence and/or structure is founded on the concept of homology and in the known proteins and annotation deposited in the databases, that is, the automatic prediction will use these information directly or indirectly. An example showing the steps of some of these databases may be built is presented in Fig. 2a, starting from DNA sequencing, generally producing complete genome sequences, to the knowledge

database, passing through identification of orthologs, clustering sequences in gene families, and automatic and manual annotations. This knowledge is then used to predict function from single proteins, complete proteomes or even metaproteomes (Fig. 2b) using many available bioinformatic tools applying different methodologies (Fig. 2c) as outlined below and detailed in Fig. 3, including commonly used tools with a simplified workflow of analysis.

2.1 Sequence-Based Methods

2.1.1 Sequence Similarity/Homology-Based Transfer

Currently, the simplest and most used method to determine protein function is based on similarity search. This is accomplished by means of similarity search programs, with BLAST (blast.ncbi.nlm.nih.gov) [25] being the most widely used form of computational function prediction methods, assigning unannotated proteins with the function of their annotated inferred homologs [10]. However, this analysis is directly dependent on databases and the annotation observed for the retrieved sequences. For that reason, when transferring function from homology inference, it is important to consider that databases contain errors, caused mainly by automatic propagation of annotation errors transferred by homology [8] and this method is, perhaps, the most sensitive to these errors. Further, the resulted database sequences, although significantly similar to query sequence, may not represent a true homolog, or may represent a paralog, instead of an ortholog, or, further, even if an ortholog was retrieved, could not present the same function (Fig. 1b). Certainly, the expansion of databases of biological sequences brought another level of problem for functional assignment. Currently, most database sequences resulting from a similarity search are hypothetical proteins with unknown function, making the analysis unfruitful and frustrating or hiding more distant-related sequences containing reliable annotation. In general, the inference of function is reliable only for very high levels of sequence identity (roughly more than 60%) [26]. An alternative to BLAST analysis is the HMMER web server (www.ebi.ac.uk/Tools/hmmer) [18] that implements protein sequence databases searches through alignments using HMM. It claims to return more correct distantly related proteins than BLAST, but HMMER search is limited to amino acid level.

Sequence similarity does not directly reflect phylogeny and may misrepresent the evolutionary structure of a phylogenetic tree [27]. As homology is an evolutionary concept, methods to infer protein function that use sequence similarity search tools (e.g., BLAST) against sequence databases should not be viewed as "homology-based," but are, instead, "similarity-based." On the other hand, the real "homology-based" methods are those exploiting phylogenetic information.

2.1.2 Protein Families and Domain Search

Domain search also include sequence similarity, but focuses on conserved motifs found in protein families. It takes into account the modular nature of the proteins and is putative more sensitive

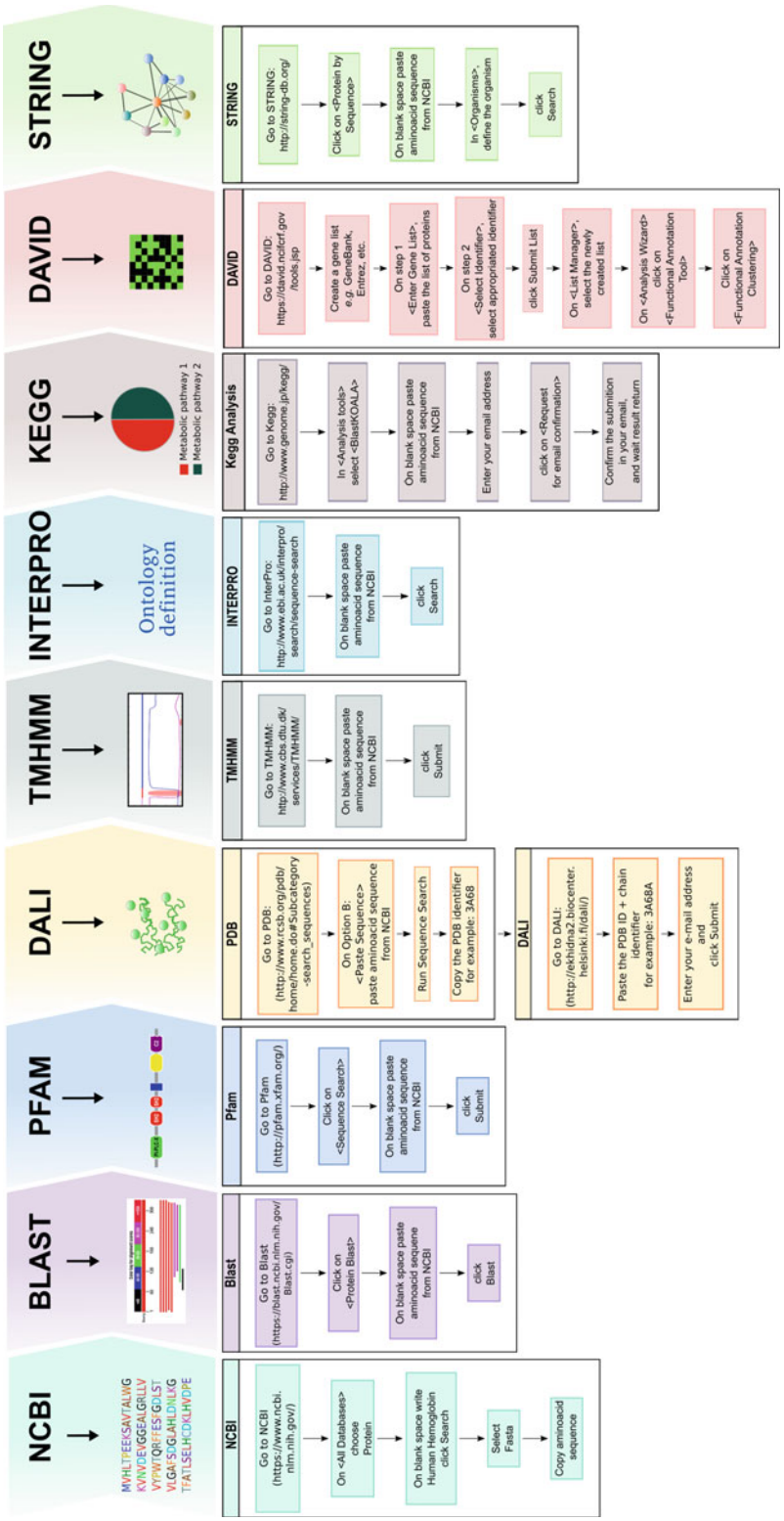

Fig. 3 Possible workflow to assign protein function using knowledge bases

because it considers only conserved regions, allowing detection of more distantly related proteins. The way used to establish motifs/domains in a protein family varies among different sources, but all start from multiple sequence alignments (MSA) of related (homologous) protein sequences in a given family. The conservation/variation in amino acids composition for each position in conserved functional regions (motifs/domains) are then extracted. The use of motifs/domains is tightly connected to protein families and can be extracted from MSA as separate single motifs/domains, multiple motifs/domains or even for the whole MSA. Conserved regions in motifs/domains are observed in MSA and described as: (a) *patterns*, a qualitative description of a motif/domain, indicating the occurrence of amino acids for each position of a motif/domain, represented through a regular expressions, as in the PROSITE database (prosite.expasy.org) [28]; (b) *profiles*, a quantitative description of a motif/domain, scoring the occurrence of each amino acid in MSA, as in Position-Specific Scoring Matrix (PSSM) used in the NCBI Conserved Domain Database (CDD) (www.ncbi.nlm.nih.gov/cdd) [29], or generating a probabilistic model using Hidden Markov Model (HMM) as in the Protein Family (Pfam) database (pfam.xfam.org) [18]; (c) *fingerprints*, groups of conserved and interrelated motifs capable to provide a signature for a particular protein family, as in the PRINTS database (www.bioinf.man.ac.uk/dbbrowser/PRINTS) [30]. These resources may be used in complementary to similarity search database analysis.

2.2 Structure-Based Methods

The function of a protein is inherently linked to its structure [31] and proteins sharing similar functions often have similar folds, a result originated from a common ancestral protein [2], the same homology concept used when comparing amino acid or nucleotide sequences. Sometimes, however, the function of one or both homologous proteins may change in the course of evolution while their folds remain largely unchanged, so in these cases the same fold may give rise to two functions [2, 26].

Methods to predict function from structure can be viewed according to the level of protein structure and specificity at which they operate, and be roughly separated in global fold similarity search and local structure definition or active site characterization [2, 31]. It should be noted, however, that not always global fold similarity correlates with functional similarity; examples include the TIM barrel fold, ferredoxin fold, and Rossmann fold global folds that are known to perform varying functions [31]. Functional assignment in these cases can be confirmed by local conservation of the residues [31]. The function of certain types of proteins is affected by a small number of residues found in a localized region of the three-dimensional structure. In enzymes, for example, the enzyme's catalytic function will be performed by a small number

of catalytic residues located in the active site [2]. Often, the specific arrangement and conformation of the residues are crucial to the performance of the function and remain strongly conserved over evolutionary time, even as the remainder of the protein's sequence and structure undergoes major changes [2]. Although global fold similarity can be used in many cases to assign a degree of functional similarity, predictions of specific biochemical or enzymatic function can be more accurately obtained from local fold similarity, i.e., in and around the protein active site [31].

Below the level of the fold come various other aspects of a protein's three-dimensional structure that may be associated with specific functions [2]. The surface of the protein, particularly its clefts and pockets, can hold important clues to function [2].

Many bioinformatics tools are available for structural function prediction. A hierarchical classification, including clusterization in homologous families, based on protein structures available in the Protein Data Bank (PDB) is presented by Class, Architecture, Topology and Homology (CATH) system (www.cathdb.info) [32] and Gene3D (gene3d.biochem.ucl.ac.uk) that uses information in CATH to predict the locations of structural domains on protein sequences from databases such as UniProtKB [33, 34]. Other methods exist for fold searching, including DALI (ekhidna.biocenter.helsinki.fi/dali_server) [35] and VAST (structure.ncbi.nlm.nih.gov/Structure/VAST) [36], which uses vector alignment of secondary structures, and CE (source.rcsb.org/jfatcatserver/ceHome.jsp) [37].

2.3 De Novo Protein Function Prediction

If an unknown protein has no significant similarity to any known protein, how is it possible to get insights about its function? In this case, computational approaches can be used to predict protein function de novo, that is, using only sequence or structure information to infer properties that are common to proteins of the same function [8]. These methods take the assumption that proteins of the same function are similarly adapted to same conditions (submitted to the same evolutionary constraints), such as pH, properties of a ligand, structural flexibility, etc. which will be reflected in their sequence and structural features [8]. Although not directly, these methods are also dependent on databases and proteins with already known function. This occurs because de novo methods generally use algorithms based on supervised learning models or statistical models, including Support Vector Machines (SVM), artificial neural networks, and Hiden Markov Model (HMM). These methods are usually less accurate than annotation transfer but are able to capture significant correlations between features and functions [8]. To do that, it needs to be "trained," that is, before scanning an amino acid sequence the models must be built from previously known proteins with the desired function or cellular

localization. These methods are largely used to establish functional residues or the subcellular localization of proteins [8].

Methods to predict functional residues assume that residues that have a similar function in different proteins are likely to possess similar physicochemical characteristics [8]. For example, residues that bind DNA share common structural and physicochemical features in most DNA-binding proteins (e.g., secondary structures, geometries, solvent accessibility, charge, hydrophobicity) [8]. There are several methods for the prediction of DNA- or metal-binding residues from sequence or structure [8].

Determining the subcellular localization of a protein helps to establish its function and can be very relevant for its experimental characterization [8]. Subcellular localization can also be predicted from similarity and motif searches if similar protein sequences with known function are available in databases, but de novo methods, instead, exploit the known correlation between amino acid composition and localization [8] and may help to even improve the knowledge about known proteins.

Many useful bioinformatics tools are available for online analysis; examples are: the Protein Subcellular Localization Prediction System (LocTree3; www.rostlab.org/services/loctree3) [38] that classifies proteins from eukaryotes, bacteria, and archaea; Balanced Subcellular Localization Predictor (BaCelLo; gpcr2.biocomp.unibo.it/bacello) [39], a predictor for the subcellular localization of proteins in eukaryotes; TargetP (www.cbs.dtu.dk/services/TargetP) [40], a predictor for eukaryotic proteins based on the presence of N-terminal signal peptide for chloroplast, mitochondrial, or secretory pathway; Subcellular Localisation Predictor (Protein Prowler; pprowler.imb.uq.edu.au) [41] determines the localization of the protein in secretory pathway, mitochondrion, or chloroplast; TMHMM (www.cbs.dtu.dk/services/TMHMM) [42] predicts transmembrane helices in protein sequences; LipoP (www.cbs.dtu.dk/services/LipoP) [43] predicts lipoproteins and signal peptides from Gram-negative bacteria protein sequences; TatP (www.cbs.dtu.dk/services/TatP) [44] predicts the presence and location of Twin-arginine signal peptide cleavage sites in bacteria.

2.4 Standard Vocabulary

Standard vocabulary on protein functional annotation provides important information to support researches on functional genomics, molecular and computational biology [4]. Schemes such as the enzyme classification system, or Enzyme Commission (EC), based on enzymatic reactions (www.chem.qmul.ac.uk/iubmb/enzyme) [45] that has been widely used in protein knowledge resources. Similarly, the Gene Ontology (GO) Consortium consists of standardized ontologies for describing gene function (www.geneontology.org) [46]. An ontology is a formal representation of knowledge by means of defined terms and its interrelationships,

allowing sequence annotation to different levels depending on the available information [46]. Both EC and GO are examples of frameworks that assign functions to groups of genes and gene products [47], creating controlled vocabulary and promoting database interoperability, but no system is directly based on protein sequences. More recently, a classification system was created for membrane transport proteins, named Transport Commission (TC), in analogy to EC system, based on the type of transport but in contrast to EC, also considers phylogenetic information based on families of homologous proteins involved (www.chem.qmul.ac.uk/iubmb/mtp) [48]. A number of other resources benefit from such controlled vocabulary, for example, the DAVID database (david.ncifcrf.gov) [49], which allows exploring functional annotation for large list of genes. EC, GO, and, more recently, TC numbers have been assigned to individual protein sequences in protein sequence databases such as UniProtKB, NCBI protein, and others. There are tools that combine standard vocabulary with similarity-based methods in predicting function from protein sequences, associating GO terms from similar proteins found in database, such as Gotcha [50] and PFP (kiharalab.org/web/pfp.php) [51], or combining different methods, including similarity and domain search, SVM and sequence derived protein features, such as CombFunc (www.sbg.bio.ic.ac.uk/~mwass/combfunc) [52] and ProtFun (www.cbs.dtu.dk/services/ProtFun) [53].

Different and complementary approaches have been applied for functional classification of proteins (and their genes) in large databases, mainly from predicted proteomes from complete genome sequences of all domains of life. These systems use bioinformatic algorithms and pipelines to generate clusters or families of protein sequences, assumed to be homologous, and classify them functionally. It is very useful in high-throughput analysis for functional classifications based on similarity search methods. Examples of those systems are The Clusters of Orthologous Groups (COG; www.ncbi.nlm.nih.gov/COG) [1], Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups (EggNOG; eggnogdb.embl.de) [54], Protein ANalysis THrough Evolutionary Relationships (PANTHER) Classification System (www.pantherdb.org) [55]. Other, special systems exist, dedicated to the classification of a more restricted group of function, for example, Carbohydrate-Active Enzymes (CAZy) database, dedicated to the families of enzymes that catalyze reactions (that degrade, modify, or create) glycosidic bonds (www.cazy.org) [56].

2.5 Systems Information

2.5.1 Genomic Context

In all organisms, the gene constitute a fundamental unit and its coded proteins tend to associate into higher levels of macromolecular complexes, biochemical pathways, and functional modules that are groups of interacting proteins acting together to accomplish a cellular process [16]. Now it is well recognized the

"modular nature" of cellular systems and this concept is considered a fundamental aspect of biological organization. Functional modules can be seen as a group of molecules acting in conjunction and interacting between them in order to perform a cellular/physiological function, with weaker connections to other functional modules [57, 58]. Frequently, functional modules show a high degree of conservation across species and may be identified in genomic associations (also linked to functional associations), such as conservation of gene order, gene/domain fusion events, and similarity of their phylogenetic profile [31, 59]. For example, the gene order is conserved in genes coding for enzymes or proteins involved in a particular metabolic pathways or cellular process, generally clustered in operons, and may serve as important clues for assigning functions if two genes retain close proximity even across large phylogenetic distances, indicating the presence of selective forces maintaining the gene organization [31]. Domain fusion is also another evolutionary event indicating functional associations in proteins, occurring when two functions are exerted by two independent proteins in one organism, but in a single protein, containing two domains in another one [31].

As an extension of genome context methods, a third indicative of functional association is the co-occurrence of genes, that is, the presence or absence of genes, known as phylogenetic profile, observed in genomes across different taxonomic groups [60]. The phylogenetic profile may be used to predict protein function by correlating the phylogenetic distribution of a query gene with that of known genes [31, 60]. The use of evolutionary information in the prediction of gene function is frequently referred as phylogenomics [61] and more elaborated methods infer function by building phylogenetic trees from homologs from known and unknown genes, generally presenting different functions assumed to rise from duplication events; the uncharacterized functions are then predicted by the phylogenetic positions relative to characterized genes [61]. Methods implemented in Orthostrapper and Function Through Evolutionary Relationships (SIFTER; sifter.berkeley.edu) [5] belong to this category.

This functional association may also be predicted via co-expression pattern in microarray analyses and/or mining literature [31]. Genome context can also be integrated with other levels of protein function information, as for example, standard vocabulary and network-based predictions. Some bioinformatics tools provide means to integrate all these levels of information, as for example, the KEGG pathway database [62] of metabolic pathway predicted from complete genome sequences, or the STRING database [63] of protein–protein interactions from different sources (including physical and functional evidences for association) and neighborhood, co-occurrence, and fusion for genes in genomic context.

2.5.2 Protein–Protein Interaction and Network-Based Prediction

One goal of modern biology is to group proteins into functional modules that act together to perform biological processes via direct and indirect interactions. The types of protein interaction within modules include physical interactions that generate protein complexes and biochemical associations [16]. Network-based predictions take advantage of these key features as gene products exhibit the tendency to associate into macromolecular complexes, biochemical pathways, and functional modules. Empirical observation shows that about 70–80% of interacting protein pairs share at least one function [24]. This observation is the rationale for methods to predict protein function using a network of protein–protein interaction, where proteins with unknown function can be assigned to the same function of known proteins interacting with them in a network. Protein–protein interaction networks can be reconstructed using proteomics, genomics, RNA expression (e.g., DNA microarrays, SGE, and RNA-seq) protein–protein interaction experiments (e.g., two-hybrid analysis, co-immunoprecipitation, and mass spectrometry), and bioinformatics approaches, which can reveal previously overlooked components and unanticipated functional associations [16, 64, 65]. The function of an unknown protein can be predicted based on its direct interactions, that is, its direct connections with known function of members observed in the network, or assisted by module, where first, groups of dense connections are identified in the network (modules), and then each module is separately annotated based on known functions of module members [66]. This approach assigns a function to an unclassified protein on the basis of function(s) present among the classified interacting proteins [24]. However, a disadvantage of this approach lies in the fact that, generally, there are few interactions observed between proteins with unknown and known functions [24].

The representation of protein–protein interactions as a network has the advantage to increase confidence levels for individual interactions and the possibility to uncover sets of protein–protein interactions that unexpectedly link diverse cellular processes or that indicate crosstalk between cellular compartments [65].

3 Final Remarks

As discussed in this chapter, the prediction of protein function is directly or indirectly dependent on proteins experimentally characterized, primary sequence and structure databases, and identification of homologous from direct sequence or structure comparison or extracted characteristics. Considering that experimentally characterized proteins are much fewer than uncharacterized proteins, and that the last continue to grow faster, automatic function prediction is the only suitable way to assign function to these “new” proteins. However, although much of these proteins with unknown

function may present homologous proteins with known function, a significant part represent orphan genes/proteins or are part of orthologous groups of unknown proteins. Further, even for unknown proteins that's function can be determined automatically, there are many reasons that makes this a complex task [3]: protein function can be studied from its molecular role to its metabolic or phenotypic effect in the whole cell; the experimental characterization of a protein is performed at a particular condition of temperature, pH, ligands concentration, etc., frequently given just partial description of its function; proteins are often multifunctional (Molecular Function and Biological Process ontologies have 30% and 60% of proteins in Swiss-Prot with more than one leaf term, respectively); annotation errors may occur due to experiment interpretation; and protein function is generally associated to gene names, difficult to predict in diverse isoforms.

Comparison of the accuracy (percentage in brackets) in predicting molecular function for experimentally characterized proteins, showed high variability in software using similarity-based methods: BLAST (75%), GeneQuiz (64%), and Gotcha (89%); and phylogeny-based methods: SIFTER (96%) and Orthostrapper (11%) [67]. A globally miss rate over 50% was found comparing the performance of Blast2GO, InterProScan, PANTHER, Pfam, and ScanProsite [68]. These results suggest the need to combine different methods when trying to predict protein functions. In a more complete survey, the performance of 54 methods for protein function prediction was evaluated by Radivojac et al. [3]. The authors established a cutoff of 60% amino acid sequence identity between an unknown and an experimentally annotated protein to be considered easy to annotate and determined its function and also observed that the overall accuracy in determining the Molecular Functional category is higher on single-domain proteins, compared to multidomain proteins [3]. The value of, at least, 60% sequence identity, and more likely closer to 80%, was also observed as required for the accurate transfer of the third level of EC classification [4].

When performing function prediction analysis important considerations should be taken into account, as outlined by Radivojac et al. [3]: (a) overall, BLAST seems ineffective at predicting functional terms in Biological Process ontology, possibly due to multiple roles played by orthologs; (b) studies have shown that correlation between sequence and function similarity is weak when applied to pairs of proteins and that domain assignments alone are not sufficient to resolve function; (c) for Molecular Function category, function prediction performance is accurate, but for Biological Process, the performance is worst; (d) methods that perform better integrate a variety of experimental evidence and weight different data appropriately for ontology terms.

A number of bioinformatics tools are available for protein function prediction and many of these tools were presented along

the text using the different methods described in this chapter. Many other useful tools are available and can be found listed and classified in reviews such as Watson et al. [2], Hawkins and Kihara [31], Friedberg [12], and Punta and Ofran ([8]—Supporting information).

References

1. Galperin MY, Makarova KS, Wolf YI, Koonin EV (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res 43: D261–D269. doi:10.1093/nar/gku1223
2. Watson JD, Laskowski RA, Thornton JM (2005) Predicting protein function from sequence and structural data. Curr Opin Struct Biol 15:275–284. doi:10.1016/j.sbi.2005.04.003
3. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Törönen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DWA, Bryson K, Jones DT, Limaye B, Inamdar H, Datta A, Manjari SK, Joshi R, Chitale M, Kihara D, Lisewski AM, Erdin S, Venner E, Lichtarge O, Rentzsch R, Yang H, Romero AE, Bhat P, Paccanaro A, Hamp T, Kaßner R, Seemayer S, Vicedo E, Schaefer C, Achten D, Auer F, Boehm A, Braun T, Hecht M, Heron M, Hönigschmid P, Hopf TA, Kaufmann S, Kiening M, Krompass D, Landerer C, Mahlich Y, Roos M, Björne J, Salakoski T, Wong A, Shatkay H, Gatzmann F, Sommer I, Wass MN, Sternberg MJE, Škunca N, Supek F, Bošnjak M, Panov P, Džeroski S, Šmuc T, Kourmpetis YAI, van Dijk ADJ, ter Braak CJF, Zhou Y, Gong Q, Dong X, Tian W, Falda M, Fontana P, Lavezzo E, Di Camillo B, Toppo S, Lan L, Djuric N, Guo Y, Vucetic S, Bairoch A, Linial M, Babbitt PC, Brenner SE, Orengo C, Rost B, Mooney SD, Friedberg I (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10:221–227. doi:10.1038/nmeth.2340
4. Clark WT, Radivojac P (2011) Analysis of protein function and its prediction from amino acid sequence. Proteins Struct Funct Bioinforma 79:2086–2096. doi:10.1002/prot.23029
5. Sahraeian SM, Luo KR, Brenner SE (2015) SIFTER search: a web server for accurate phylogeny-based protein function prediction. Nucleic Acids Res 43:W141–W147. doi:10.1093/nar/gkv461
6. Galperin MY, Koonin EV (2010) From complete genome sequence to "complete" understanding? Trends Biotechnol 28:398–406. doi:10.1016/j.tibtech.2010.05.006
7. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y (1998) Predicting function: from genes to genomes and back. J Mol Biol 283:707–725
8. Punta M, Ofran Y (2008) The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS Comput Biol 4: e1000160
9. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12:85–94
10. Sleator RD (2012) Prediction of protein functions. In: Kaufmann M, Klinger C (eds) Functional genomics. Springer, New York, NY, pp 15–24
11. Sleator RD, Walsh P (2010) An overview of in silico protein function prediction. Arch Microbiol 192:151–155. doi:10.1007/s00203-010-0549-9
12. Friedberg I (2006) Automated protein function prediction – the genomic challenge. Brief Bioinform 7:225–242. doi:10.1093/bib/bbl004
13. Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8:995–1005. doi:10.1038/nrm2281
14. Khan I, Chen Y, Dong T, Hong X, Takeuchi R, Mori H, Kihara D (2014) Genome-scale identification and characterization of moonlighting proteins. Biol Direct. doi:10.1186/s13062-014-0030-9
15. Jeffery CJ (1999) Moonlighting proteins. Trends Biochem Sci 24:8–11
16. Hu P, Janga SC, Babu M, Díaz-Mejía JJ, Butland G, Yang W, Pogoutse O, Guo X, Phanse S, Wong P, Chandran S, Christopoulos C, Nazarians-Armavil A, Nasseri NK, Musso G, Ali M, Nazemof N, Eroukova V, Golshani A, Paccanaro A, Greenblatt JF, Moreno-Hagelsieb G, Emili A (2009) Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. PLoS Biol 7:e1000096. doi:10.1371/journal.pbio.1000096

17. Madupu R, Richter A, Dodson RJ, Brinkac L, Harkins D, Durkin S, Shrivastava S, Sutton G, Haft D (2012) CharProtDB: a database of experimentally characterized protein annotations. Nucleic Acids Res 40:D237–D241. doi:10.1093/nar/gkr1133
18. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR (2015) HMMER web server: 2015 update. Nucleic Acids Res 43:W30–W38. doi:10.1093/nar/gkv397
19. Goodacre NF, Gerloff DL, Uetz P (2014) Protein domains of unknown function are essential in bacteria. mBio 5:e00744-13. doi:10.1128/mBio.00744-13
20. Bateman A, Coggill P, Finn RD (2010) DUFs: families in search of function. Acta Crystallogr Sect F Struct Biol Cryst Commun 66:1148–1152. doi:10.1107/S1744309110001685
21. Theißen G (2002) Orthology: secret life of genes. Nature 415:741–741. doi:10.1038/415741a
22. Zakon HH (2002) Convergent evolution on the molecular level. Brain Behav Evol 59:250–261
23. Doolittle RF (1994) Convergent evolution: the need to be explicit. Trends Biochem Sci 19:15–18. doi:10.1016/0968-0004(94)90167-8
24. Vazquez A, Flammini A, Maritan A, Vespignani A (2003) Global protein function prediction from protein–protein interaction networks. Nat Biotechnol 21:697–700
25. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. doi:10.1186/1471-2105-10-421
26. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y (2003) Automatic prediction of protein function. Cell Mol Life Sci CMLS 60:2637–2650. doi:10.1007/s00018-003-3114-8
27. Engelhardt BE, Jordan MI, Srouji JR, Brenner SE (2011) Genome-scale phylogenetic function annotation of large and diverse protein families. Genome Res 21:1969–1980. doi:10.1101/gr.104687.109
28. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I (2013) New and continuing developments at PROSITE. Nucleic Acids Res 41:D344–D347. doi:10.1093/nar/gks1067
29. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH (2015) CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222–D226. doi:10.1093/nar/gku1221
30. Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, Roma-Mateo C, Theodosiou A, Mitchell AL (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource – its status in 2012. Database 2012:bas019. doi:10.1093/database/bas019
31. Hawkins T, Kihara D (2007) Function prediction of uncharacterized proteins. J Bioinforma Comput Biol 5:1–30
32. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG, Lehtinen S, Studer RA, Thornton J, Orengo CA (2015) CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res 43:D376–D381. doi:10.1093/nar/gku947
33. Lam SD, Dawson NL, Das S, Sillitoe I, Ashford P, Lee D, Lehtinen S, Orengo CA, Lees JG (2016) Gene3D: expanding the utility of domain assignments. Nucleic Acids Res 44:D404–D409. doi:10.1093/nar/gkv1231
34. Yeats C, Lees J, Carter P, Sillitoe I, Orengo C (2011) The Gene3D web services: a platform for identifying, annotating and comparing structural domains in protein sequences. Nucleic Acids Res 39:W546–W550. doi:10.1093/nar/gkr438
35. Holm L, Rosenstrom P (2010) Dali server: conservation mapping in 3D. Nucleic Acids Res 38:W545–W549. doi:10.1093/nar/gkq366
36. Gibrat JF, Madej T, Bryant SH (1996) Surprising similarities in structure comparison. Curr Opin Struct Biol 6:377–385
37. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11:739–747
38. Goldberg T, Hecht M, Hamp T, Karl T, Yachdav G, Ahmed N, Altermann U, Angerer P, Ansorge S, Balasz K, Bernhofer M, Betz A, Cizmadija L, Do KT, Gerke J, Greil R, Joerdens V, Hastreiter M, Hembach K, Herzog M, Kalemanov M, Kluge M, Meier A, Nasir H, Neumaier U, Prade V, Reeb J, Sorokoumov A, Troshani I, Vorberg S, Waldraff S, Zierer J, Nielsen H, Rost B (2014) LocTree3 prediction of localization. Nucleic Acids Res 42:W350–W355. doi:10.1093/nar/gku396
39. Pierleoni A, Martelli PL, Fariselli P, Casadio R (2006) BaCelLo: a balanced subcellular localization predictor. Bioinformatics 22:e408–e416. doi:10.1093/bioinformatics/btl222
40. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular

localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005–1016. doi:10.1006/jmbi.2000.3903

41. Boden M, Hawkins J (2005) Prediction of subcellular localization using sequence-biased recurrent networks. Bioinformatics 21:2279–2286. doi:10.1093/bioinformatics/bti372
42. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes 11 Edited by F. Cohen. J Mol Biol 305:567–580. doi:10.1006/jmbi.2000.4315
43. Juncker AS, Willenbrock H, von Heijne G, Brunak S, Nielsen H, Krogh A (2003) Prediction of lipoprotein signal peptides in gram-negative bacteria. Protein Sci 12:1652–1662. doi:10.1110/ps.0303703
44. Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S (2005) Prediction of twin-arginine signal peptides. BMC Bioinformatics 6:167
45. du Plessis L, Skunca N, Dessimoz C (2011) The what, where, how and why of gene ontology – a primer for bioinformaticians. Brief Bioinform 12:723–735. doi:10.1093/bib/bbr002
46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29
47. Lesk AM (2010) Introduction to protein science: architecture, function, and genomics, 2nd edn. Oxford University Press, Oxford
48. Saier MH (2006) TCDB: the transporter classification database for membrane transport protein analyses and information. Nucleic Acids Res 34:D181–D186. doi:10.1093/nar/gkj001
49. Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57. doi:10.1038/nprot.2008.211
50. Martin DM, Berriman M, Barton GJ (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 5:178. doi:10.1186/1471-2105-5-178
51. 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. doi:10.1110/ps.062153506
52. Wass MN, Barton G, Sternberg MJE (2012) CombFunc: predicting protein function using heterogeneous data sources. Nucleic Acids Res 40:W466–W470. doi:10.1093/nar/gks489
53. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Stærfeldt HH, Rapacki K, Workman C, Andersen CAF, Knudsen S, Krogh A, Valencia A, Brunak S (2002) Prediction of human protein function from post-translational modifications and localization features. J Mol Biol 319:1257–1265. doi:10.1016/S0022-2836(02)00379-0
54. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293. doi:10.1093/nar/gkv1248
55. Mi H (2004) The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 33:D284–D288. doi:10.1093/nar/gki078
56. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42:D490–D495. doi:10.1093/nar/gkt1178
57. Wagner GP, Pavlicev M, Cheverud JM (2007) The road to modularity. Nat Rev Genet 8:921–931. doi:10.1038/nrg2267
58. Pereira-Leal JB, Levy ED, Teichmann SA (2006) The origins and evolution of functional modules: lessons from protein complexes. Philos Trans R Soc B Biol Sci 361:507–517. doi:10.1098/rstb.2005.1807
59. Osterman A, Overbeek R (2003) Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol 7:238–251. doi:10.1016/S1367-5931(03)00027-9
60. Kensche PR, van Noort V, Dutilh BE, Huynen MA (2008) Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. J R Soc Interface 5:151–170. doi:10.1098/rsif.2007.1047
61. Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8:163–167. doi:10.1101/gr.8.3.163
62. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–D462. doi:10.1093/nar/gkv1070
63. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic

M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43:D447–D452. doi:10.1093/nar/gku1003

64. Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast 18:523–531
65. Mayer ML, Hieter P (2000) Protein networks—built by association. Nat Biotechnol 18:1242–1243. doi:10.1038/82342
66. Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol. doi:10.1038/msb4100129
67. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE (2005) Protein molecular function prediction by Bayesian Phylogenomics. PLoS Comput Biol 1:e45. doi:10.1371/journal.pcbi.0010045
68. Rodrigues BN, Steffens MBR, Raittz RT, Santos-Weiss ICR, Marchaukoski JN (2015) Quantitative assessment of protein function prediction programs. Genet Mol Res 14:17555–17566. doi:10.4238/2015.December.21.28