A Survey of Computational Methods for Protein Function Prediction

Amarda Shehu, Daniel Barbará, and Kevin Molloy

Abstract Rapid advances in high-throughout genome sequencing technologies have resulted in millions of protein-encoding gene sequences with no functional characterization. Automated protein function annotation or prediction is a prime problem for computational methods to tackle in the post-genomic era of big molecular data. While recent community-driven experiments demonstrate that the accuracy of function prediction methods has significantly improved, challenges remain. The latter are related to the different sources of data exploited to predict function, as well as different choices in representing and integrating heterogeneous data. Current methods predict function from a protein's sequence, often in the context of evolutionary relationships, from a protein's three-dimensional structure or specific patterns in the structure, from neighbors in a protein–protein interaction network, from microarray data, or a combination of these different types of data. Here we review these methods and the state of protein function prediction, emphasizing recent algorithmic developments, remaining challenges, and prospects for future research.

Keywords Computational biology • Protein function prediction • Algorithms • Machine learning • Homology

A. Shehu (✉)
Department of Computer Science, Department of Bioengineering, George Mason University, Fairfax, VA 22030, USA
e-mail: amarda@gmu.edu

D. Barbará
Department of Computer Science, George Mason University, Fairfax, VA 22030, USA
e-mail: dbarbara@gmu.edu

K. Molloy
LAAS-CNRS, 7, avenue du Colonel Roche, 31077 Toulouse, France
e-mail: kmolloy@laas.fr

K.-C. Wong (ed.), *Big Data Analytics in Genomics*,
DOI 10.1007/978-3-319-41279-5_7

1 Introduction

Molecular biology now finds itself in the era of big data. The focus of the field on high-throughout, automated wet-laboratory protocols has resulted in a vast amount of gene sequence, expression, interactions, and protein structure data [212]. In particular, due to the increasingly fast pace with which whole genomes can be sequenced, we are now faced with millions of protein products for which no functional information is readily available [39, 198]. The December 2015 release of the Universal Protein (UniProt) database [68] contains a little over 55.2 million sequences, less than 1 % of which have reliable and detailed annotations.

The gap between unannotated and annotated gene/protein sequences has exceeded two orders of magnitude. Fundamental information is currently missing for 40 % of the protein sequences deposited in the National Center for Biotechnology Information (NCBI) database; around 32 % of the protein sequences in the comprehensive UniProtKB database are currently labeled “unknown.” The missing information includes coarse-grained, low-resolution information such as where protein products are expressed, meta-resolution information, such as what chemical pathways proteins participate in the living cell, and high-resolution information, such as what molecular partners a protein recognizes and binds to directly in the cell.

Getting at what proteins do in the living cell is central to our efforts to understand biology, as proteins are ubiquitous macromolecules [4] involved in virtually every cellular process, from cell growth, maintenance, proliferation, to apoptosis [5]. Understanding what a protein does in the cell is also central to our ability to understand and treat disease [351]. Moreover, computer-aided drug design (CADD) often begins with identifying a protein target whose activity in the diseased cell needs to be altered or regulated via binding compounds to cure or treat disease [322].

Given the exponential increase in the number of protein sequences with no functional characterization and the central role of proteins to human biology and health, predicting where a protein acts in the cell and exactly what it does is a central question to address in molecular biology. Originally, this question was only investigated in the wet laboratory and on a small set of target genes or proteins. Wet-laboratory approaches that elucidate the role of a protein in the cell include gene knockout, targeted mutations, inhibition of gene expressions, mass spectrometry, and RNAi [5].

Gene knockout, targeted mutations, and inhibition of gene expression methods demand considerable effort and time and can only handle one protein product or gene at a time [306]; in other words, these are low-throughput methods. Higher-throughput wet-laboratory annotation initiatives, such as the European Functional Analysis Network [264] have also proven unable to keep up with the pace of whole genome sequencing. In particular, wet-laboratory experiments that use mass spectrometry or RNAi are found to yield biased and less specific information about protein function than the low-throughput methods [309].

Human experts known as biocurators also often peruse published wet-laboratory studies to provide functional information on a protein [38, 308, 309]. For instance, the popular UniProtKB database, which is the central hub for the collection of functional information on protein sequences, consists of two sections: SWISS-PROT, which contains manual and reviewed annotations (less than 1 % of protein sequences in UniProt are annotated in SWISS-PROT), and TrEMBL, which contains computational annotations yet to be validated by experts [34]. About 30–40 % of computational and manual annotations contain errors [308]. The rapid rise of the CRISPR/Cas technology [67, 140], which can edit specific genes and do so rather efficiently [303], proves promising to observe the phenotypic effect of deliberately introduced gene variants [134, 210, 286], but applications for large-scale function prediction are under-pursued at the moment.

In light of the increasing gap between the amount of protein sequence data and the amount of functionally annotated proteins, computational approaches seem poised to tackle protein function prediction and narrow this gap. Before venturing into a comprehensive description of such methods, which is the subject of this review, it is important to formulate exactly what one means by protein function. One can find different definitions in literature, because the function of a protein can be described at different degrees of detail. Information on the cellular localization of a protein can provide important clues towards the processes in which a protein is involved but is not sufficient in itself. Describing function from a physiological aspect entails knowing the biological processes in which a protein participates. From a phenotypical aspect, one is more concerned about the disease or disorders induced by a misbehaving protein. To capture all these different aspects and aid computational approaches, various classification schemes have been proposed. By now, the most broadly accepted and utilized scheme is the Gene Ontology (GO) scheme originally proposed in [12]. GO is a hierarchical description of protein function that describes three different aspects, each one increasing the level of detail.

- **Cellular component** describes the component or anatomical structure in a cell where a gene product operates. Examples include the rough endoplasmic reticulumn, nucleus, ribosome, proteasome, and more.
- **Biological process** captures the physiological description of protein function and allows specifying the processes in which a gene product participates in the cell. A process is defined as a series of events or molecular functions, and examples include membrane fusion, cellular component organization, macromolecular complex assembly, and various distribution processes.
- **Molecular function** is different from the biological processes in which a gene product is involved and instead captures at a finer level of resolution what a protein does in the cell, such as transporting molecules around, binding to molecules, holding molecular systems together, changing systems from one state to another. Examples include ligand binding, catalysis, conformational switching, and more.

Computational methods for protein function prediction are diverse, particularly when one considers methods that limit themselves to prediction of specific aspects of protein function. The focus of this review is on methods that aim to provide GO annotations of protein products. These methods can be organized in distinct categories based on the type of data they employ to predict the GO annotation of an uncharacterized protein.

The first category of computational methods is comprised of methods that make predictions based on observations that sequence similarity is a good indicator of functional similarity. Such methods were among the first to be employed for automatic annotations, and they are often the first tools employed in this regard. They are summarized in Sect. 2. Recent advancements in sequence-based methods concern expanding their applicability beyond proteins of very high sequence similarity, known as close homologs, to remote homologs. These methods consider additional information such as genomic context and evolutionary relationships and are described in Sects. 3 and 4.

Another category is comprised of methods that use more than sequence information and employ information on the three-dimensional, biologically active structure of a protein. This information is often difficult to obtain in the wet laboratory and not available on many proteins. However, advances in structure resolving techniques, both in the wet and dry laboratories, are allowing the application of such methods for protein function prediction. Structure-based methods are mainly distinguished by the representations they choose of protein structure and the amount of protein structure they exploit. These methods are described in Sect. 5.

Yet another category is comprised of methods that employ information on known interactions of a protein product as encapsulated in protein–protein interaction networks. This category is rich in machine learning methods and is the subject of Sect. 6. Methods that exploit gene expression data are summarized in Sect. 7.

Currently, the best-performing methods are those that are enriched with additional information on sequence, structure, and gene expression data, resulting in a category of methods known as hybrid methods, which we describe in Sect. 8. Another category of methods, described in Sect. 9, exclusively mine biomedical literature to annotate query proteins.

Methods for automated function prediction are now evaluated and tested in community-driven experiments and global initiatives, such as the Enzyme Function Initiative (EFI), the COMputational BRidges to EXperiments (COMBREX) initiative, and the Critical Assessment of Function Annotation (CAFA) community-driven experiment. In particular, CAFA is becoming the main venue to objectively compare function prediction methods to one another and highlight the state of the art in automated function prediction [287]. Evaluation is done in two rounds. In CAFA1, several thousands of unannotated query sequences are provided to participants. About 48,298 targets from 18 species were provided in 2014. Participants submit predicted GO terms, and predictions are then evaluated according to community-agreed metrics, such as the top-20, threshold measure, and the maximum F1 score over all recall-precision pairs obtained with the threshold measure, also known as the Fmax score [124] and others.

In the second round of CAFA, CAFA2, twice as many queries are released. For instance, 100,816 target protein sequences from 27 species were provided to participants in 2014. Evaluation of predicted GO terms from different labs is then held as a special interest group meeting at the Intelligent Systems in Molecular Biology (ISMB) conference. Annual reviews are released tracking progress, challenges, and the state of the art in automated function prediction. We summarize the latest review of function prediction methods in the context of their performance in CAFA to conclude our survey of these methods. In particular, the survey concludes with a critical summary of the state of protein function prediction, remaining challenges, and prospects for future research.

2 Sequence-Based Methods for Function Prediction

Sequence-based methods transfer onto an uncharacterized target protein sequence the functional annotation of a characterized protein sequence with high sequence similarity to the target. Some of the earliest efforts in bioinformatics focused on understanding the relationship between sequence, structure, and function similarity. This was made possible by the advent of standardized sequence formats and sequence comparison tools, such as FASTA [278], and fast sequence alignment and comparison algorithms based on dynamic programming, such as BLAST and PSI-BLAST [8]. In addition to BLAST, other well-known sequence alignment tools now include PROSITE [18, 143] and PFAM [327, 328].

The comparison of two sequences aims to determine an evolutionary relationship and infer whether the sequences under comparison share a common ancestor; that is, if they are homologs. One cannot infer shared ancestry, thus homology, based on sequence similarity alone. For instance, high sequence similarity might occur because of convergent evolution; when considering shorter sequences, high similarity may occur because of chance. Two sequences can be similar but not homologous. Therefore, function cannot be realizably transferred even if sequence similarity is high.

While early bioinformatics efforts (some of which are summarized below) indeed transferred function between highly similar sequences, later efforts, aware of convergent evolution, focused on integrating additional data beyond sequence. It is worth noting that early efforts focused on understanding when sequence similarity is high enough, in the absence of convergent evolution, to infer function similarity. By now it has been observed that prediction accuracy suffers when the threshold is set to anything less than 30 % sequence similarity [159]; however, a more comprehensive understanding has emerged that shows that, as long as homology is established, even 10 % sequence identity can allow extracting information on function (for instance, remote homologs can have very low sequence identities due to early branching points in evolution). On the other hand, even 30 % sequence identity between two non-homologous sequences can be misleading [184].

Sequence-based methods have been shown to have limited applicability for some of the reasons listed above. In response, more sophisticated sequence-based methods have been developed. These methods either enhance the basic framework, where the functional unit is still the comparison of two protein sequences in their entirety, or pursue complementary frameworks of comparing subsequences or physico-chemical properties extracted from two given protein sequences. Based on this distinction, sequence-based methods can be organized in three categories:

- **Sequence-based methods**: methods in this category rely on the comparison of a query protein sequence to functionally annotated sequences in a database. More sophisticated methods pursue a probabilistic setting, incorporate data from various sources, and even pursue unsupervised learning frameworks to improve the accuracy and confidence of annotations.
- **Subsequence-based methods**: methods in this category realize that only a subset of the amino acids in a protein comprise the site onto which molecular partners bind. Methods in this category mainly differ in what subsequences are considered, domains or shorter subsequences known as motifs.
- **Feature-based methods**: these methods aim to extract more information from a given protein sequence than what is directly available in the identity of amino acids. By additionally encoding physico-chemical properties of amino acids, these methods construct features for all or a subset of a sequence and pursue functional annotation in a machine learning setting.
- **Ensemble-based methods**: these methods combine the above three approaches via the concept of ensemble classifiers in machine learning.

2.1 Sequence-Based Methods for Functional Annotation

Several directions have been investigated to improve the performance and/or extend the applicability of sequence-based methods. We review representative methods below.

2.1.1 Sequence Alignment-Based Methods

The accuracy of transferring functional annotations was investigated in [133], where it was shown on annotation transfer among enzymes that transfer was only reliable when sequence similarity was high. Many subsequent studies trying to determine the sequence identity threshold below which functional annotation transfer was not reliable resulted in the recognition of a twilight zone 25–30 % sequence identity [299]. While basic sequence-based methods are not reliable when sequence identities are 25 % or less, sophisticated methods can handle the twilight zone. The ConFunc method proposed in [368] operates in this range. PSI-BLAST is used to align a query sequence with annotated sequences. The sequences returned by PSI-BLAST are then split into sub-alignments according to the sequences' GO annotations. Conserved residues are then identified within each GO-term sub-alignment, and a position-specific scoring matrices (PSSM) profile is constructed for each sub-alignment. The query sequence is scored against the PSSMs of all sub-alignments, and these scores are then used to calculate expectation values for the GO annotations corresponding to each sub-alignment. Prediction is made based on careful filtering of the different GO annotations based on their expectation values. ConFunc is shown to outperform both BLAST and PSI-BLAST. On a large testing set of query sequences with known homologs with sequence identities in the twilight zone, ConFunc's recall is six times greater than BLAST.

A related method, GoFDR, is proposed in [116], which processes the PSI-BLAST query-sequence based on multiple sequence alignment (MSA). For each GO term of the homologs in the MSA, GFDR identifies functionally discriminating residues (FDR) specific to the GO term. The query sequence is then scored using a position-specific scoring matrix constructed for the FDRs alone. The raw score is converted into a probability based on a score-probability table prepared over training sequences. GoFDR outperforms three sequence-based methods for predicting GO terms, PFP [129], GOtcha [236], and ConFunc [368], and is ranked as the top method in the preliminary evaluation report in CAFA2.

2.1.2 Probabilistic Whole-Sequence Annotation Transfer

One direction pursues sequence-based functional annotation in a probabilistic setting. The approach proposed in [207] assumes that a protein can only belong to a functional class if its BLAST score distribution with members of the class is the same as that of these members with one another. A univariate and multivariate probabilistic scheme are investigated. The univariate scheme makes predictions based on the total score of the query protein, by assigning to it a probability of belonging to each functional class. This can lead to ambiguous results, as the query can have similar scores with different functional classes. For this reason, the univariate scheme is extended to a multivariate one by constructing a vector of BLAST scores of the query with all classes. The vector is compared to the distribution of each class. Evaluation shows that the approach reaches an accuracy above 90 % [207]. However, the evaluation is performed on enzymes, where sequences are more strongly correlated with function than on other proteins. In addition, this probabilistic approach performs well on the most specific GO level due to the ambiguity with comparisons to the less specific GO levels (such as cellular location).

2.1.3 Integration of Data Sources in Whole-Sequence Annotation Transfer

Another way to improve whole-sequence methods is to integrate additional data sources. The GOtcha method in [236], for instance, organizes the annotations of sequences similar to a query into a set of GO-like directed acyclic graphs. A P-score is calculated based on the frequency of occurrence of respective annotations and BLAST E-values of the corresponding matches. The P-score estimates the confidence attached to the annotation of the query sequence with that term, and a threshold value for the P-score allows extracting a final set of annotations. Evaluation of this approach on the *Drosophila melanogaster* genome showed that the results were more sensitive and specific than those obtained with the baseline approach.

2.1.4 Unsupervised Learning in Whole-Sequence Annotation Transfer

Work in [1, 380] pursues an unsupervised learning approach. Sequences similar to a query sequence are identified via BLAST. All pairwise sequence similarities in the set, including the query, are stored in a similarity matrix. The latter is employed for clustering the set of sequences. The annotation of the query sequence is then based not on individual high-similarity sequences but on the cluster of sequences to which the query sequence belongs. In [380], progressive single-linkage clustering and text information analysis are employed to assign GO terms to the query sequence. In [1], the sequence similarity space is encoded in a graph, and the normalized cut clustering algorithm is used to identify groups of sequences that are closely related to the query sequence. In [289], the space of annotated sequences is first organized via hierarchical clustering according to functional and evolutionary relationships. The function of the query sequence is then predicted based on the position of the query in the tree. The approach employed in these methods uses whole-sequence comparisons only as an intermediate step and maps a query protein to a cluster or a level of a hierarchy. This approach is shown to perform well and be more robust to errors in individual entries [1, 289, 380].

A specific subgroup of methods that fall in the same category do not directly address function prediction but rather construct an informative organization of protein sequences into functional groups. The objective is to extract from the groups rules and features that can then be utilized by other methods in a machine learning setting. Specifically, protein sequences are clustered into functional groups based on their evolutionary relationships, structural properties (these can be structural classes based on secondary structure content, folds, or even structural motifs), or subsequences. Manual curators can be employed to provide such clustering, but our focus here is on computational methods. Iterative clustering is proposed in [314, 382]. The method in [94] encodes proteins into a graph with edges encoding pairwise similarity, and then applies Markov Clustering to group known homologs from different species (orthologs) [211]. In [385], clusters of related proteins are identified by analyzing strongly connected sets of vertices in the graph. In [242], pairwise sequence comparisons are used to organize proteins into pre-families, which are then further divided into homogeneous clusters based on the topology of the similarity graph. Spectral clustering is employed in [257] to infer protein families. A heuristic approach is proposed in [10] to improve the performance of these graph-based clustering methods related to the choice of the similarity threshold determining whether two proteins should be connected by an edge or not. Others use hierarchical clustering [56, 226, 305, 342, 350]. Partitioning clustering algorithms have been shown recently to outperform all these other clustering methods [99].

2.1.5 Supervised Learning and Generative Models in Whole-Sequence Annotation Transfer

The most successful methods for function prediction rely on supervised learning or generative models. Methods proposed in [66, 83, 150, 213] address the detection of remote homologs by employing sequence alignment profiles to train hidden Markov models (HMMs) or pairwise sequence similarities to train support vector machines (SVMs) or neural networks (NNs). Specifically, the FANN-GO method proposed in [66] aligns a query sequence to a database of annotated sequences to calculate the i-score proposed by the GOtcha method [236] that the query is associated with a specific functional term. The scores are then fed to an ensemble of multi-output neural networks trained to predict the probability of a sequence associated with each function term. The FANN-GO method is shown able to model dependencies between functional terms and outperforms the GOtcha method [236].

2.2 Subsequence-Based Methods for Functional Annotation

Subsequence-based methods are motivated by a deeper understanding of the role of protein sequence in recognition events. In particular, only a subset of the amino acids that comprise a protein chain assist in the sticky interactions with other molecules. This subset is typically comprised of few amino acids. In addition, contiguous, long segments of a protein chain may fold independently to form a domain. Multi-domain proteins employ different domains to interact with different molecules and thus enrich their molecular functions [314]. Subsequence-based methods are organized into two categories:

- **Domain-based methods**: It has long been recognized that a multi-domain protein's array of functions is due to different molecular functions of its

(continued)

domains [314]. Domain-based methods seek to identify all the domains in a protein in order to compile its array of molecular functions.

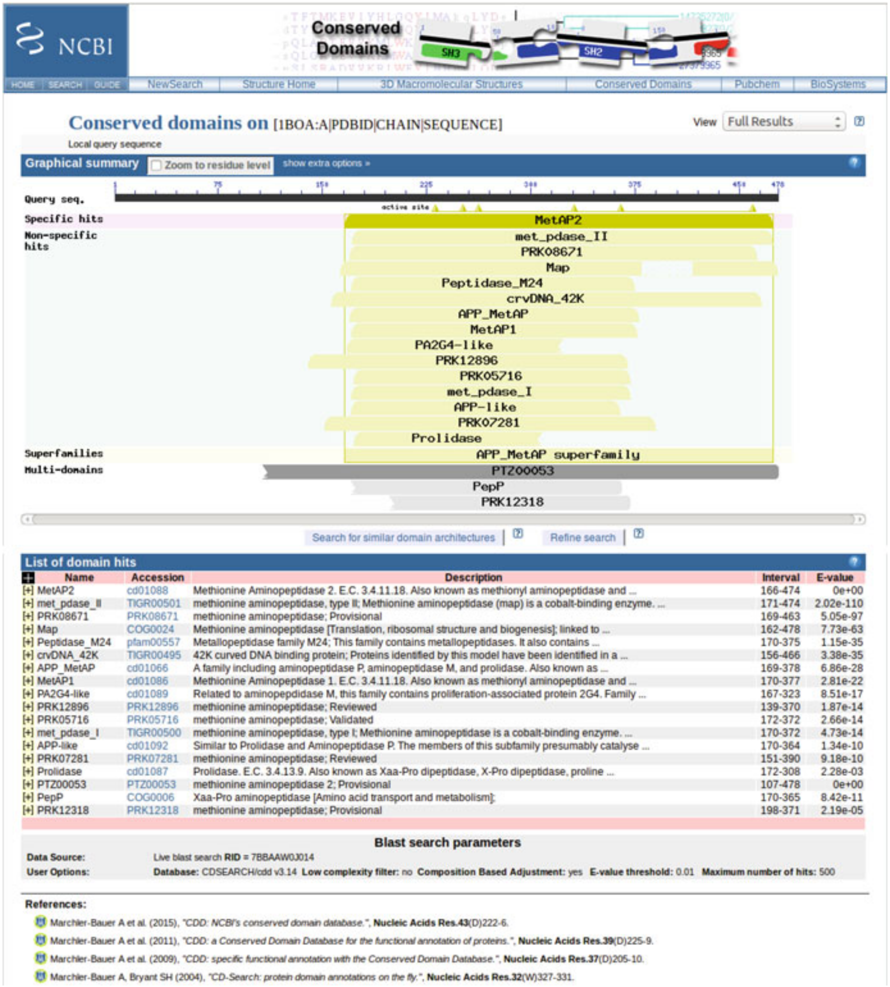
- **Motif-based methods**: Functional sites in a protein that are employed to recognize and bind ligands, DNA, RNA, and other proteins are comprised of only a subset of the amino acids comprising a protein's polypeptide chain. Since functional sites are under higher evolutionary pressure to be conserved, a way of identifying functional sites on proteins is through detection of evolutionary-conserved (sequence) subsequences. Subsequences that are conserved among protein sequences belonging to a family are referred to as *motifs* [35]. These methods detect motifs in a protein sequence and use these motifs as signatures of specific functional classes [141].

Motif-based methods are rich in machine learning techniques. In contrast, domain detection methods focus more on integrating biological insight. We review domain-based methods first, and then devote the rest of the description of subsequence-based methods to the machine learning strategies employed to detect motifs for functional annotation.

2.2.1 Domain-Based Methods

A protein domain is an independent evolutionary and functional unit of a protein that folds independently of the rest of the protein where it is contained. A transcribed exon is referred to as a module, and a domain may be comprised of several modules [346]. More than 80 % of known domains are about 50–150 amino acids long, but exceptionally long domains of more than 800 can be found, as well. Several small domains of 30 amino acids or less are also reported. At least two-thirds of mammalian proteins have more than one domain. Only multi-cellular eukaryotic organisms have a significant proportion of proteins with repeating domains [344].

A domain is an independently folding unit of a protein. So, a domain is a structural unit that can be found in multiple protein contexts. Biologists usually break up large proteins into domains based on a process that involves analysis of sequence, structure, and domain-specific expertise. Protein domains can be found in several databases, such as ProDom [314] and the Conserved Domain database (CDD) [230]. These two databases describe domains at the sequence level. A query protein sequence is aligned to the deposited domains, and BLAST E-values are employed to determine the domains present in the query. In particular, CDD is the protein classification component of NCBI's Entrez query and retrieval system by which one can identify conserved domains in query protein sequences. An illustration of the results returned by the CD-Search tool (which stands for conserved domain search) in CDD is shown in Fig. 1.



Conserved domains on [1BOA:A|PDBID|CHAIN|SEQUENCE]

Local query sequence

List of domain hits

Name	Accession	Description	Interval	E-value
[+] MetAP2	cd01088	Methionine Aminopeptidase 2. E.C. 3.4.11.18. Also known as methionyl aminopeptidase and ...	166-474	0e+00
[+] met_pdase_II	TIGR00501	methionine aminopeptidase, type II; Methionine aminopeptidase (map) is a cobalt-binding enzyme. ...	171-474	2.02e-110
[+] PRK08671	PRK08671	methionine aminopeptidase; Provisional	169-463	5.05e-97
[+] Map	COG0024	Methionine aminopeptidase [Translation, ribosomal structure and biogenesis]; linked to ...	162-478	7.73e-63
[+] Peptidase_M24	pfam00557	Metallopeptidase family M24; This family contains metallopeptidases. It also contains ...	170-375	1.15e-35
[+] crvDNA_42K	TIGR00495	42K curved DNA binding protein; Proteins identified by this model have been identified in a ...	156-466	3.38e-35
[+] APP_MetAP	cd01066	A family including aminopeptidase P, aminopeptidase M, and prolidase. Also known as ...	169-378	6.86e-28
[+] MetAP1	cd01086	Methionine Aminopeptidase 1. E.C. 3.4.11.18. Also known as methionyl aminopeptidase and ...	170-377	2.81e-22
[+] PA2G4-like	cd01089	Related to aminopepdidase M, this family contains proliferation-associated protein 2G4. Family ...	167-323	8.51e-17
[+] PRK12896	PRK12896	methionine aminopeptidase; Reviewed	139-370	1.87e-14
[+] PRK05716	PRK05716	methionine aminopeptidase; Validated	172-372	2.66e-14
[+] met_pdase_I	TIGR00500	methionine aminopeptidase, type I; Methionine aminopeptidase is a cobalt-binding enzyme. ...	170-372	4.73e-14
[+] APP-like	cd01092	Similar to Prolidase and Aminopeptidase P. The members of this subfamily presumably catalyse ...	170-364	1.34e-10
[+] PRK07281	PRK07281	methionine aminopeptidase; Reviewed	151-390	9.18e-10
[+] Prolidase	cd01087	Prolidase. E.C. 3.4.13.9. Also known as Xaa-Pro dipeptidase, X-Pro dipeptidase, proline ...	172-308	2.28e-03
[+] PTZ00053	PTZ00053	methionine aminopeptidase 2; Provisional	107-478	0e+00
[+] PepP	COG0006	Xaa-Pro aminopeptidase [Amino acid transport and metabolism];	170-365	8.42e-11
[+] PRK12318	PRK12318	methionine aminopeptidase; Provisional	198-371	2.19e-05

Blast search parameters

Data Source: Live blast search **RID** = 7BBAAW0J014

User Options: **Database:** CDSEARCH/cdd v3.14 **Low complexity filter:** no **Composition Based Adjustment:** yes **E-value threshold:** 0.01 **Maximum number of hits:** 500

References:

Marchler-Bauer A et al. (2015), *"CDD: NCBI's conserved domain database."*, **Nucleic Acids Res.43**(D)222-6.

Marchler-Bauer A et al. (2011), *"CDD: a Conserved Domain Database for the functional annotation of proteins."*, **Nucleic Acids Res.39**(D)225-9.

Marchler-Bauer A et al. (2009), *"CDD: specific functional annotation with the Conserved Domain Database."*, **Nucleic Acids Res.37**(D)205-10.

Marchler-Bauer A, Bryant SH (2004), *"CD-Search: protein domain annotations on the fly."*, **Nucleic Acids Res.32**(W)327-331.

Fig. 1 Output obtained by the CD-Search tool on the human methionine aminopeptidase 2 protein. Domain hits are provided visually and ranked from non-specific (high E-values) to specific (low E-values). Details on each of the domain hits are listed in the *bottom panel*

Other databases also store domains and provide more than sequence and functional information. The CATH [266, 277] and SCOP [251] databases contain structural information, as well. These databases provide hierarchical classifications of protein structures. For instance, CATH breaks down a query protein structure into Class (C), then Architecture (A), followed by Topology (T), and then Homology(H). SCOP breaks down a query protein structure into fold, then superfamily, then family. Both CATH and SCOP are discussed in greater detail in Sect. 5, which discusses structure-based methods for functional annotation. It is worth noting, however, that due to different working definitions of domains and different techniques used to determine what makes a domain, these two databases contain a different number of domains. CATH contains at the moment about 65,000 domains, whereas SCOP contains about 110,800 domains. SCOP defines a domain as an evolutionary unit.

Manual curation of domains provides at the very least sequence and functional information that has been exploited early on to functionally annotate new proteins. In [311], a query protein sequence is aligned to sequences of domains extracted from ProDom and CDD, and rules for function assignment are based on BLAST E-values. Application to a set of 4357 manually curated human proteins results in a recall of 81 % and a precision of 74 %; on data sets from other organisms, precision and recall are decidedly lower, at about 50 %.

A machine learning approach is proposed in [48], where domains extracted from the SBASE library of protein domains [359] are employed as attributes. Sequence alignment is used to determine whether a domain is present or not in a given protein sequence, allowing a protein sequence to be represented as a binary vector recording domain membership. k-nearest neighbor (kNN) and SVM were then trained on functionally annotated protein sequences. kNN is reported to outperform SVM on 13 functional classes obtained from the MIPS database [48].

Work in [281] follows a slightly different approach. Instead of using domains for function annotation, constant-length statistically significant sequence patterns known as promotifs [343] are used instead. Correlations found between SWISS-PROT keywords assigned to the sequences and positions of promotifs are then used to establish rules for function annotation. Precision on a set of PROSITE [144] protein sequences is reported similar to that obtained from work in [311]. A low sensitivity of 50 % is reported. A similarly-low sensitivity is obtained by application of decision trees classifiers to recognize PFAM domains [329] in [131]

Work on domain detection is very rich in computational biology, and methods often include more than sequence information to reliably identify domains in a given protein. One reason for domain-based functional prediction is that relying upon sequence to detect domains falls short. Sequence information is found to be insufficient for identifying structural domains in a protein, because the same structure can be assumed from highly divergent sequences of less than 30 % sequence identity. Structural-based domain detection is more reliable to identify the domains in a protein, but this requires knowledge of a protein's folded structure, which is not readily available for many unannotated protein sequences. A comprehensive review of structural-based domain detection methods can be found in [356].

Independent of the technique used to identify domains present in a given protein, the problem of how to use this information to assign function(s) to an unannotated protein remains open. This is due to the fact that a unit designated to be a domain may not meet the definition of a domain as an independent evolutionary and functional unit of a protein that folds independently of the rest of the protein. Hence, when the criteria for what makes a unit a domain are indeed relaxed, the relationship between a so-called domain and its function is not clear. In addition, domain domain interactions need to be detected, as they may give rise to more complex molecular functions. A review of the current understanding of the domain function relationship can be found in [346, 361]. As such understanding improves, sophisticated machine learning methods that take into account context rather than membership and domain-domain interactions are expected to improve the state of domain-based functional annotation.

2.2.2 Motif-Based Methods

The Motif Alignment and Search Tool (MAST) [17] was one of the first to be used for motif-based function annotation. MAST employs the MEME algorithm [16] to construct groups of probabilistic motifs. MAST combines these motifs to construct protein profiles for protein families of interest. MAST successively estimates the significance of the match of a query protein sequence to a particular family model as the product of the p-values of each motif match score. This measure is then used to select the family of the query sequence. MAST's average classification accuracy (ROC50) over 72 distinct queries is shown to be above 0.95.

While MAST is a statistical approach, other methods based on unsupervised and supervised learning approaches have been proposed for motif-based functional annotation, as well. The algorithm in [217] implements unsupervised learning, but relies on SPLASH rather than MEME to detect motifs. Other algorithms, some of which implement supervised learning, employ motifs extracted from motif libraries and databases.

The SPLASH algorithm proposed in [49] is the first to define motifs as sparse amino acid patterns that match repeatedly in a set of protein sequences. This definition is amenable to computational approaches for motif detection. The SPLASH algorithm is a deterministic pattern discovery algorithm that is extremely efficient and can be used in a parallel setting to systematically and exhaustively identify conserved subsequences in protein family sets. The algorithm is employed in the first motif-based method for function annotation [217] to extract motifs. The resulting set of motifs are enriched via a HMM. Proteins are then clustered based on motif membership in a top-down clustering algorithm reminiscent of the construction of a decision tree. The levels of the tree correspond to the motifs, ranked from most to least significant, and proteins are divided at each level/node of the tree based on whether they contain the corresponding motif or not. In this manner, the leaves contain sets of proteins that contain all motifs that together define the signature of a particular functional family. The method performs well, achieving a classification rate between 57 % and 72 % on the exceptionally challenging and sequence-diverse G Protein-Coupled Receptor (GPCR) superfamily.

Instead of the unsupervised method in [217], work in [365] pursues a supervised learning setting to classify proteins based on motif membership. The latter is used to represent each protein sequence as a binary vector, with each entry indicating the presence or absence of a particular motif; motifs are extracted from the PROSITE database [144]. A decision tree is then learned on manually curated protein families extracted from the MEROPS database [290]. Classification accuracy is reported to be significantly better than the clustering approach in [49].

An NN-based method is proposed in [33], using two different types of motifs, class-independent and class-dependent motifs. The former refer to motifs extracted from the training dataset, and the latter refer to motifs extracted for each known functional class. About 30 motifs from each class are extracted with the MEME algorithm [16]. Class-dependent motifs are found to confer the best classification performance to the NN. Application of the NN on the GPCR superfamily yields better classification performance than MAST [17]. However, on this challenging superfamily, a naive Bayes classifier with a χ^2 feature selection algorithm is shown to obtain the best performance [60]. It is worth noting that the special focus on GPCRs is due to their central role in drug design; about 60 % of the approved drugs target some member of the GPCR family [107].

Work in [27] conducts a comprehensive machine learning study for motif-based function annotation. A motif kernel uses the occurrence count of each motif in a protein sequence as a similarity measure between the motif and the sequence. Classification is carried out via an SVM. Issues such as feature selection and multi-class classification are also investigated. An evaluation of different strategies for feature selection and classification schemes is conducted on a set of enzymes. The best performance is obtained when using the RFE method [123] for feature selection, many one-against-the rest classifiers [295], and counting the multiple classes of a protein as a single class. In addition, SVM is found to perform better than kNN. This result is further confirmed in [191], where motif-based SVMs perform best on the classification of enzymes.

2.3 Feature-Based Methods for Functional Annotation

Many of the methods already described can be viewed as feature-based methods, since they employ term identify or term frequency representations of protein sequences in unsupervised or supervised learning settings, with terms being domains, motifs, modules, promotifs, or even genomes. However, here we describe a category of sequence-based methods that expand their treatment of a protein sequence beyond the identity of amino acids to physico-chemical properties of amino acids. At their core, these methods expand a protein sequence into a vector of physico-chemical properties or attributes of amino acids in the sequence, otherwise referred to as a feature vector. Once such a transformation is made, standard classification techniques can be used. The most popular ones for feature-based functional annotations of protein sequences are the SVM, NN, kNN, and naive Bayes classifiers.

Supervised learning has also been employed to improve and extend the applicability of whole-sequence annotation transfer in the twilight zone of low sequence similarity. The main approach is to organize BLAST or PSI-BLAST results into a positive set, which contains sequences with high similarity to the query sequence, and the negative set, which contains all other sequences. The sets are then used to train a classifier to discriminate sequences in the twilight zone.

One of the earliest methods, PROCANS [376], employs a three-layered NN and represented protein sequences as binary vectors. The attribute for each amino acid in a sequence records the conservation of the amino acid at that position in an MSA and valued 1 when at a position with 50–100 % identity and 0 otherwise. Later improvements to this algorithm employ n-grams as features. In [374], n-gram counts are used to construct feature vectors. In [373, 375], the order of n-grams in a sequence is found to be important, and positions of n-grams in a sequence are added to the feature vector. This collection of works, however, is limited to enzymes, where there are stronger correlations between sequence and function.

The PRED-CLASS method in [274] also employs an NN but restricts its focus to three specific protein classes. The NN contains three levels, classifying transmembrane proteins at the first level, fibrous proteins at the second level, and globular proteins at the third level. Different features are employed at different levels. Compositional features are employed at the first level, as transmembrane proteins have specific subsequence signature; 30 features record the composition of a sequence in all 20 amino acids and 10 different groupings of residues with common structural and physico-chemical properties. 30 features are employed at the second and third level, but these features correspond to the 30 highest intensity for periodicities detected for each residue or each of the 10 considered groups of residues. PRED-CLASS is reported to correctly classify 96 % of 387 proteins.

Work in [172] expands the applicability of feature-based methods. Propositional data mining and inductive logic programming are employed in [172] on binary feature vectors constructed from protein sequences. The features correspond to sets of characteristics found in a significant fraction of the protein sequences. A C4.5 decision tree classifier is trained to predict rules from these features. An encouraging prediction accuracy of 65 % is obtained on ORFs of *Mycobacterium tuberculosis* and *Escherichia coli*.

An extension of the above method is described in [173], where a comparative study is conducted to determine representations of protein sequences that improve function prediction. Three types of representations are investigated, sequence-based attributes, phylogeny-based attributes, and structure-based attributes. Sequence-based attributes (SEQ) are based on a sequence's composition of single and pairs of amino acids. Phylogeny-based attributes (SIM) are computed from the sequences returned from a PSI-BLAST search of a given ORF sequence. The attributes capture, via a first-order language such as Datalog, the distribution of sequences, their evolutionary distance from the ORF, the phylogenetic relationship, as well as keywords describing the sequences; the latter could be easily computed from a sequence, such as the presence of a membrane/trans-membrane binding sequence. Structure-based attributes (STR) are computed from secondary structure segments predicted for a sequence from the Prof program [270]. Pair combinations of these representations, as well as a combination of all of them, are also investigated. Evaluation on ORFs in *E. coli* demonstrates that SIMs confer the highest performance. This important result was one of the first indications that evolutionary information is powerful and perhaps more informative than sequence. Section 4 describes methods that exploit evolutionary history for function prediction.

Work in [153] proposes the now popular ProtFun method, which trains a set of NNs on carefully constructed attribute-value pairs from a training data set. The attributes are compiled via different tools that provide information on post-translational modifications, such as N- and O-glycosylation, phosphorylation, cleavage of N-terminal signal peptides, and other modifications and sorting events that a protein is subjected to before performing its function. An extensive evaluation of ProtFun on 5500 human proteins from the TrEMBL database, with annotations assigned based on SWISS-PROT keywords via the EUCLID system, demonstrates that ProtFun has a high sensitivity of 90 % and a low false positive rate of 10 % on some functional categories. Similar high performance is reported in [154], where ProtFun is applied to predict GO categories for human proteins.

SVMs and kNNs have recently shown to be superior for feature-based function classification. The SVM-Prot method proposed in [47] employs physico-chemical properties of amino acids to represent a protein sequence as a 5-entry feature vector. Five attributes are considered for each amino acid, normalized van der Waals volume, polarity, charge, and surface tension. The attributes are averaged over all amino acids in a sequence, resulting in a five-dimensional feature vector. Binary classification is carried out for each protein family, with proteins in the family comprising the positive data set and all other proteins the negative data set. Evaluation is carried out on annotated proteins extracted from several databases, and reported classification accuracies are in 69.1–99.6 % range. In [125], SVM-Prot is reported to obtain an accuracy of 71.4 % on a testing data set of 49 plant proteins.

Work in [86] attempts to address some of the issues with predicting GO categories for protein sequences. The Classification in a Hierarchy Under Gene Ontology (CHUGO) system is proposed, which recognizes that assignment of a protein to a particular GO node, immediately assigns the protein to all ancestors of the node in the GO hierarchy. Therefore, labels in CHUGO are not specific GO categories, but GO subgraphs in the GO hierarchy. Since CHUGO trains a separate binary classifier for each GO node, effectively an ensemble of classifiers are used for a protein, as a protein can belong to multiple classes at a particular level in the GO hierarchy. While these ideas are merit-worthy, the performance of CHUGO is not higher than sequence-based methods for function prediction.

Work in [240] uses the HMM proposed in [164] to construct profiles for protein families and classify novel proteins via profile comparisons. Families are identified via a single-link hierarchical clustering algorithm of 256,413 manually annotated proteins. The families are available via the PANTHER/Lib library provided as part of this work, and families are carefully indexed according to their GO-based ontology terms documented in the PANTHER/X to permit fast GO annotation transfer on a query sequence. PANTHER v.8.0 now has 82 complete genomes organized into gene families and subfamilies and has evolved to providing not only gene function, ontology, but also pathways and statistical analysis tools. The PANTHER system has emerged as a highly popular tool that enables browsing and query of gene functions, and a large-scale gene functional analysis has recently been reported via the PANTHER classification system [239].

Work on effective feature representations that encode the low-level constraints that function places on sequence is expected to continue. In fact, feature engineering is considered the most promising direction in sequence-based methods for function prediction. Recent work in [262] focuses on this direction and provides the community with hundreds of features of high biological interpretability and are shown promising in predicting subcellular localization, structural classes, and unique functional properties (such as thermophilic and nucleic acid binding).

2.4 Ensemble-Based Methods

State-of-the-art sequence-based approaches typically employ ensemble techniques that combine the three categories of sequence-based methods described above. For instance, work in [170] proposes two ensemble methods, the consensus (CONS) method and the frequent pattern mining (FPM) method. Each method combines GO predictions from PFP [129, 130], ESG [63], PSI-BLAST [8], PFAM [329], FFPred [225], and HHblits [293]. Each of the ensemble methods in [170] improved performance over the individual methods in the CAFA1 and CAFA2 categories of the CAFA competition.

The GOPred method proposed in [304] combines heterogeneous classifiers that cover the three main sequence-, subsequence-, and feature-based approaches to improve GO annotations. Positive and negative training data sets prepared for each of the 300 GO molecular function terms are subjected to three classification methods, each representative of the three main approaches: BLAST k-nearest neighbor (BLAST kNN), the subsequence profile map (SPMap), and the peptide statistics with SVMs (Pepstats-SVM) method. Four classifier combination techniques are investigated: majority voting, mean, weighted mean, and addition. Evaluation in [304] demonstrates that the weighted mean classifier combination technique, which assigns different weights to the classifiers depending on their discriminative power for a specific functional term, achieves the best performance in 279 of 300 classifiers.

2.5 State of Sequence-Based Function Prediction

An interesting observation based on evaluation of different methods for function prediction at CAFA has been that sequence-based methods set the bar high for function prediction [124]. While any current function prediction method ought to outperform sequence-based methods, in practice, improvements are small, as sequence-based methods, when implemented correctly, do very well at CAFA. Such are the conclusions of a survey of such methods in 2013 [124]. In particular, it is observed that sequence-based methods can perform at the top-20 provided precise details of the implementation are followed. For instance, score normalization across targets and poor choices for values of free parameters can lead to lower performance. Careful implementations of sequence-based methods can lead to high performance on the CAFA top-20 and threshold measures.

Many challenges remain regarding sequence-based function prediction. One particular challenge regards the detection of remote homologs, which are homologous sequences with less than 25 % sequence identity thought to make up about 25 % of all sequenced proteins. It is worth noting that a particular group of machine learning methods address the problem of remote homology detection. These methods are both sequence-, subsequence-based, and profile-based and predominantly employ SVMs [26, 189, 288] or HMMs [151, 164, 218–220, 249, 323]. A comparative review of some of these methods can be found in [104]. These methods are typically evaluated not directly on function prediction but instead on reproducing the SCOP superfamily classification [9], which is considered the gold standard due to its manual curation [296] (though recently a different picture is emerging of errors in SCOP and increasing reliability of CATH classification due to improvements of machine learning methods for automated structural classification). It remains to be seen whether these methods, by improving the detection of remote homologs can further improve automated functional annotation of proteins.

3 Genomic Context Methods for Function Prediction

Genomic context-based methods can be the only viable approach in cases of query proteins with novel sequences, and for which interacting molecular partners have yet to be discovered. These methods are predicated on the knowledge that location of the gene encoding a query protein is important information that can be exploited for function prediction. These methods fall into two main categories:

- **Gene neighborhood- or gene-order based methods**: these methods operate under the hypothesis that two proteins with corresponding genes located in proximity of each other in multiple genomes are expected to interact functionally.
- **Gene fusion-based methods**: these methods operate under the hypothesis that pairs or sets of genes identified in a genome that are merged into a single gene in another genome are functionally related.

3.1 Gene Neighborhood- and Gene Order-Based Methods

The hypothesis that gene proximity in a genome implies functional interactions between the proteins they encode is supported by the concept of an operon, which contains one or more genes that are transcribed as a unit in mRNA. The concept of the neighborhood was originally exploited in [71], which found that 75 % of neighboring genes were known to interact physically, with the rest representing potentially novel interactions. Work in [271, 272] inferred functional coupling between genes in 24 genomes, which additionally employed the concept of a pair of close bidirectional best hits (PCBBH); neighboring genes in a genome G1 with neighboring orthologs in a genome G2. PCBBH entries were scored based on the evolutionary distance between two genomes, and the scores were employed to report as predictions those PCBBH entries with scores above a pre-defined threshold. Work in [185] improves upon the idea of PCBBHs by addressing the issue that gene proximity is not sufficient in itself to infer functional coupling. Additional constraints beyond neighboring orthologs, such as proximity of transcription start sites and opposite direction of transcription are enforced in [185] to infer functional coupling.

The SNAPper method in [179, 180] relies on the construction of a similarity-neighborhood graph (SN-graph). Vertices in the graph are the genes in a given set of genomes. Edges connect vertices corresponding to orthologs or neighboring genes. The notion of an SN-cycle is employed, which is hypothesized to preferentially join functionally related gene products that participate in the same biochemical or regulatory process. Evaluation demonstrates that SNAPper is more effective at reconstructing metabolic pathways than directly predicting functional annotations.

3.2 Gene Fusion Methods

Gene fusion was first proposed in [233], which hypothesized that if two genes are separate in one genome but are merged or fused into a single gene in another genome, then these genes are expected to be functionally related. There is strong biological reasoning to support this hypothesis. Gene fusion reduces entropy of disassociation, indicating that genes that encode for two domains of one protein in an earlier organism evolve into independent genes in a descendant organism [95]. The hypothesis is also supported at a structural level, since it has been observed that protein–protein interfaces are highly similar to domain–domain interfaces in a multi-domain protein [349].

The effectiveness of the basic idea of gene fusion was demonstrated in [233] by analyzing 6809 pairs of non-homologous genes in the *E. coli* genome. A significant fraction of these pairs were found to have been reported physically or functionally. The basic premise of gene fusion was validated at a large scale in [383], which applies the method to 30 microbial genomes and reports an average sensitivity of

72 % and an average specificity of 90 % on predicted functional links. Another study on 24 genomes reports similarly high performance [93]. The method in [232] expands upon the basic gene fusion approach by replacing orthology for the broader concept of homology. An association scoring function based on the hypergeometric distribution measures the probability of the chance occurrence of a given number of fusion events between a pair of genes. The log of this association score is found to have a linear correlation with functional similarity.

Gene neighborhood- and gene fusion-based methods have been shown in various large-scale studies [147] to yield interactions that are functionally meaningful, such as direct physical interactions, co-membership in a protein complex, co-presence in metabolic or non-metabolic pathways, or other biological processes. Many databases, such as Phydbac [91] and Phybac [89], now store functional associations detected by gene neighborhood- and gene fusion-based methods, as well as phylogenetic profiles. In general, however, methods based on the notion of gene neighborhoods are more accurate at finding functional links than methods based on gene fusion and phylogenetic profiles (described in Sect. 4).

Several systematic studies of genome context methods are now available. For instance, work in [101] carries out a thorough comparison of many different methods on data from several organisms. Several conclusions are drawn. For instance, the study finds gene fusion methods to generally perform the worst, and gene neighbor methods to outperform phylogenetic profile methods by as much as 40 % in sensitivity on most organisms.

4 Phylogenomics-Based Methods for Function Prediction

Phylogenomics-based methods expand genomic context and exploit evolutionary relationships between organisms to detect functional similarities between genes. These methods fall into three categories:

- **Phylogenetic profile-based methods**: these methods encode the presence or absence of a gene across genomes in a binary vector referred to as the phylogenetic profile. The underlying hypothesis is that two genes with similar phylogeny profiles will also be functionally similar.
- **Phylogenetic tree-based methods**: these methods exploit the concept of a phylogeny tree, which encodes evolutionary relationships and distances between organisms in a tree. The pattern of evolution of a set of proteins, as present in phylogeny trees, can be exploited in a machine learning setting to detect functional similarity.

(continued)

- **Phylogeny hybrid methods**: Recent machine learning methods combine the information present in phylogenetic profiles and trees.

4.1 Phylogenetic Profile-Based Methods

The operating hypothesis for phylogenetic profile-based methods for function prediction is that proteins that participate in the same pathway or molecular complex in the cell are under pressure to evolve together to preserve their role in the cell. Hence, the comparison of phylogenetic profiles, which store in a binary vector the presence or absence of a particular gene in a genome is expected to be valuable for function comparison.

This hypothesis is tested in [279]. Phylogenetic profiles are constructed from 16 genomes of different organisms. Three *E. coli* proteins are used as the test data set to verify that indeed proteins with profiles different at most one position are functionally related. SWISS-PROT annotations are employed for the verification. Similar results are derived from the EcoCyc database of metabolic pathways [168]. In addition, a comparative study in [214] demonstrates that comparison of phylogenetic profiles is more accurate in predicting function than comparison of whole protein sequences. The study additionally shows that function prediction accuracy improves if more genomes are included to construct phylogenetic profiles.

An important extension of phylogenetic profile-based methods is also proposed in [214] regarding the scenario of redundant genes in an organism that are eventually lost. Such genes can be detected by allowing for complementary phylogenetic profiles, as shown in [214] for DNA-directed DNA polymerases, DNA repair proteins, and isomerases. The PhylProM database computes these results and other phylogenetic data [345].

Work on phylogenetic profile-based methods has mainly proceeded in three directions.

The first direction concerns investigating distance functions for comparing two phylogenetic profiles. In [377], Hamming Distance, Pearson's correlation coefficient, and Mutual Information (MI) are compared on the ability to determine co-membership in a metabolic pathways in KEGG [162], and MI is found to confer the best performance.

The second direction concerns investigating different representations of phylogenetic profiles beyond the binary vector. Work in [88] proposes real-valued profiles, where entries record not just membership in a genome but instead record the normalized BLAST score denoting the best match for a protein in a genome. This is effectively a relaxation of the phylogenetic profile to cases where an exact match for a gene cannot be found in a genome. Cosine similarity is used to identify the neighborhood of a profile, and annotation is carried out by finding the statistically dominant class of the MultiFun database [313]. Comparison with the original profile-based method in [279] shows that the relaxation idea provides better performance. It is worth noting that the Phydbac database [89] includes profile-based annotations obtained from real-valued profiles. The Phydbac2 database [90] by the same team of researchers strengthens the annotation procedure by combining predictions based on the genomic context methods described in Sect. 3.

Work in [29] pursues both above directions of what information to encode in profiles and how to compare phylogenetic profiles. Two modifications are proposed. Based on prior studies showing that including more genomes improves accuracy, partially complete genomes are proposed to be included in constructing the phylogenetic profile of a gene. Second, the distance function is proposed to take into consideration both the number of genomes and the evolutionary history of the unannotated gene. Comparing two profiles constructed from a large number of genomes is assigned greater significance than when comparing profiles constructed from fewer genomes. The farther a genome where the query gene is found is in evolutionary history from the genome containing a gene, the higher the weight of the corresponding entry in distance calculations. Evaluation of these ideas on detecting co-occurrence in the KEGG database for two distinct test sets does not show significant performance improvements over earlier work; however, performance is impacted more by the number of genomes than differently weighting profile entries in the distance function [29].

The third direction of research on phylogenetic profile-based methods pursues the combination of information extracted from phylogenetic profiles with that extracted from gene neighborhoods and gene fusion from genomic context-based methods. For instance, the PLEX method in [73] proceeds in iterations. In the first, genes with similar phylogenetic profiles to a query gene are first identified. These genes are then used as queries to identify possibly more genes with similar profiles in another iteration, and this process continues until no new genes are identified. The predicted functional links are then combined with those obtained from gene neighbor and gene fusion links. This approach has been shown capable of reconstructing the important urease enzyme complex and the isoprenoid biosynthesis pathway in *M. tuberculosis*. Another method employs similar ideas [392] but constructs a single, tandem phylogenetic profile for pairs of neighboring genes in a genome. The profile records whether a pair of neighboring genes in a genome are also neighboring in other genomes. This is in itself an interesting genomic context-based extension of the concept of phylogenetic profiles. Pairs with similar profiles are collected, and the functional coherence is tested with respect to functional categories in the Clusters of Orthologuous Groups (COG) proteins database [340]. Comparison of purity and the Jaccard coefficient show that pair profiles confer better performance than single profiles. Comparison of different distance functions shows that MI is more powerful.

Due to the incredibly intuitive relationship between phylogenetic profiles and functional annotations, methods exploiting phylogenetic profiles have not had to investigate sophisticated techniques from machine learning beyond a comparison of distance functions. However, several directions can be investigated. For instance, other similarity measures are known to be more powerful than MI in machine learning, such as odd ratio, Yule's Q and Yule's Y, as well as Piatetsky-Shapiro and collective strength [335]. Association rule mining based on market basket analysis [2] has also not been investigated, though it may identify meaningful frequent patterns from a matrix of phylogenetic profiles. These lines of investigation may further improve the performance of phylogenetic profile-based methods for function prediction.

4.2 Phylogenetic Tree-Based Methods

Phylogenetic tree-based methods are also prime for investigating different ideas and techniques from machine learning; however, current research on phylogenetic tree-based functional annotation is scarce, primarily for two reasons. First, it is decidedly more difficult and intricate to compare trees. Second, phylogenetic trees are constructed via algorithms themselves and so are approximations of the actual, unknown evolutionary tree for the organisms under consideration.

Work in [85] demonstrated in 1998 how phylogenetic trees could be used for predicting function. Specifically, the homologs of a query protein can be identified via homology-based methods. The query protein and its homologs can then be embedded in a phylogenetic tree via tree reconstruction algorithms, such as PHYLIP [100]. Gene duplication and speciation events can then be identified in the tree and be used to assign function to the query protein. Work in [85] shows that this approach can be more reliable than homology-based methods when there are variations to the rate of functional change and gene duplication events and changes to the function of homologs during evolution. The promise of phylogenetic trees to improve function prediction over homology-based methods is also demonstrated in [80].

Based on this foundational work, later methods have focused on exploiting phylogenetic trees for identifying domain–domain and protein–protein interactions [276], and for training generative and discriminative models of molecular function [92, 285, 321]. In [276], the similarity between two phylogenetic trees is interpreted as an indication of coordinated evolution and similar evolutionary pressure to members of a given molecular complex. Rather than comparing phylogenetic trees, the method in [276] compares the distance matrices that are used to build phylogenetic trees. Earlier work in [112] introduces such a similarity measure by measuring the linear correlation coefficient between all sets of pairwise distances in the tree.

In [285], a HMM is constructed at each parent node of the phylogenetic tree by using the multiple sequence alignment of the reconstructed sequences of the child nodes. A score is associated with each node, and the query protein is assigned the class whose tree scores the highest. A very high accuracy of 99 % is achieved for a set of 1749 GPCRs. In [321], a multi-step strategy is proposed and disseminated via the GTREE software. The strategy begins with identifying homologs of the query, constructing a multiple sequence alignment of the query and its homologs, using the alignment to eventually construct a phylogenetic tree, identifying the high support subtrees, integrating additional experimental data, differentiation of orthologs and paralogs to infer molecular function.

Work in [92] proposes the SIFTER method, which is based on probabilistic graphical models. The seminal idea is that a reconciled phylogenetic tree can be considered a probabilistic graphical model if transition probabilities are associated with its edges. So, a transition probability function is assigned for the transfer of molecular function from a parent to a node. Then, standard propagation algorithms are used to compute the posterior probability of a node being assigned a certain molecular function. An extensive comparative evaluation shows that the method is superior to sequence-based methods and other phylogenetic tree-based methods. A very high precision is reported with complete coverage on 100 Pfam families supplemented with GO annotations. In 2005, SIFTER beat other methods, such as BLAST, GeneQuiz, GOtcha, GOtcha-exp, and Orthostrapper. Specifically, SIFTER achieves 96 % prediction accuracy against a gold standard dataset of 28 manually annotated proteins in the AMP/denosine deaminase Pfam family. SIFTER performs better than BLAST, GeneQuiz, GOtcha, GOtcha-exp (GOtcha transferring only experimental GO annotations), and Orthostrapper, which achieve 75 %, 64 %, 89 %, 79 %, and 11 % prediction accuracy, respectively.

4.3 Hybrid Phylogenetic Profile and Tree Methods

Since the phylogenetic profile of a protein sequence and phylogenetic trees encode different evolutionary knowledge, they can be combined. The method in [358] uses SVMs to learn protein function from phylogenetic profiles. The phylogenetic tree is used to define a tree kernel that calculates profile similarity. Performance comparison between the SVM with the tree kernel vs. the SVM with a linear kernel on the genes of *Saccharomyces cerevisiae* shows that the tree kernel confers better classification performance. The method in [358] is adapted in [256] to use real-valued rather than binary phylogenetic profiles. The elements of a phylogenetic profile are obtained by a post-order traversal of the phylogenetic tree; the internal nodes of the tree are assigned scores that are averages of the children scores. An SVM with a polynomial kernel is trained on the resulting real-valued profiles and shown to outperform the original method in [358].

5 Structure-Based Methods for Function Prediction

Despite increasingly sophisticated sequence-based methods for function prediction, research has shown that sequence is not under a strong selective pressure to preserve function. A prime example of this is the presence of remote homologs, which were identified as early as the 1960s, when Perutz and colleagues showed through structural alignment that myoglobin and hemoglobin had similar structures but indeed different sequences [282]. By now many studies have shown that the correlation between sequence and structure is stronger than that between sequence and function [79, 138, 369]. In particular, structure is under more evolutionary pressure than sequence, and structure-based methods effectively cast a wider net at detecting functional similarity based on structural similarity. Some studies suggest that utilizing both sequence-to-structure and structure-to-function relationships may allow for more accurate function prediction [102].

The goal in structure-based methods for function prediction is to detect a level of similarity between two given protein structures, which in terms allows for the transfer of functional annotations from one protein to another. Similarity can be detected by comparing the two structures in their entirety or only in part. This distinction allows organizing structure-based methods into the following categories:

- **Whole Structure-based methods**: these methods identify similar protein structure utilizing a distance metric to transfer functional annotations from a query structure. Most methods rely on a structural alignment which is prohibitive at a large scale, when comparing a query structure to structures in a database. The high computational cost can be addressed by filter-based methods, which attempt to reduce the number of structures of relevance for comparison to the query.
- **Substructure-based methods**: Similar to the subsequence-based methods discussed earlier, substructure-based methods consider that only a portion of the structure may contain the binding sites critical for binding with molecular partners. A query protein is effectively searched for substructures known to exist between functionally related proteins.

5.1 Whole Structure-Based Methods

The growth in the size of structural databases such as the Protein Data Bank (PDB) [28] is enabling the transfer of functional annotations between structurally similar proteins. Similar to sequence comparisons, structural alignments are used to enable the comparison of proteins of unequal size. Alignment methods strive to maximize the number of residues in the alignment while minimizing some distance or similarity measure. While many structural alignments methods have been proposed, it remains unclear which method provides the most biologically significant alignment [320, 331]. For a given alignment, an optimal superimposition of two proteins is commonly performed using the Kabsch method [160], which finds the rotations require to minimize the root-mean-squared-deviation (RMSD), between the atoms identified in the alignment. RMSD is a popular distance measure for protein structures. Another popular distance measure is GDT_TS [387], which measures the maximum percentage of C_α atoms that can be aligned within a set of difference tolerances (typically 1Å, 2Å, 4Å, and 8Å).

Below, we discuss two main subclasses of whole sequence-based methods. The first uses superimposition and alignments. While these methods are accurate, they are also computationally demanding. The second subclass of methods lower computational demands by reducing or eliminating the need for alignments and superimpositions.

5.1.1 Alignment-Based Whole Structure-Based Methods

Some of the most popular alignment-based structure comparison methods are SALIGN [36], SSM [175], MAMMOTH [268], CE [319], STRUCTAL/LSQMAN [176, 206, 332] SSAP [267], VAST [108], SARF2 [6], and DALI [136]. These methods have to address three main issues: how to represent a protein structure so as to facilitate comparison, how to efficiently explore the space of possible alignments in search of the optimal alignment or near-optimal ones, and how to score a given alignment and determine its statistical significance.

Structural alignment is typically cast as a multi-objective optimization problem, where the best alignment is obtained from maximizing the number of amino acids included in the alignment and in tandem improving a structural alignment score measuring structural similarity. The choice of the alignment technique and the particular score employed are independent of each other, but many methods couple them together.

Computing the alignment is computationally demanding. Some methods compute alignments that are optimal with respect to a similarity score, while others compute approximate alignments in return for computational expediency. The secondary-structure mapping (SSM) method presented in [175] aligns two given proteins using their secondary structure units and employs a scoring function similar to VAST [108]. VAST normalizes aligned structures by incorporating the chances of randomly finding these structures in the PDB (much like term frequency inverse document frequency (TF-IDF) in text mining). VAST is now part of the NCBI's structure computational services.

DALI [136] utilizes internal distance matrix representations of protein structures and a sequence score between matched residues to perform the alignment. An initial alignment is generated and a Monte Carlo scheme is used to find better alignments. An integer linear programming approach utilizing the DALI scoring function is proposed in [370] to compute the optimal alignment. The contact map overlap (CMO) scoring method is proposed by Godzik and Skolnick in [111], which also uses internal distances to maximize the number of contacts in the alignment. The TM-Align method and the TM-Score are proposed in [390], where coordinate superimposition and dynamic programming are coupled with an optimized scoring matrix.

Many of the alignment-based methods are available to the community via web servers. Web servers that allow users to employ and compare many of these methods are also available. For instance, the CSA web server for comprehensive structural alignments [372] provides the community with access to both exact and heuristic alignment methods, as well as many different scoring functions.

Alignment-based methods have been the focus of periodic reviews [177, 182, 234]. In particular, in [177, 182], many of these alignment-based methods are tested on a benchmark set of almost 3000 structures and found to perform well but to carry a large computational cost that becomes prohibitive when the goal is to find the structural neighbor of a protein in a database of 50,000 or more structures.

Recently, alternative approaches have been proposed that do not rely on alignment. For instance, work in [371] utilizes a metric based on contact map overlaps. Another set of methods avoid the alignment and superimposition requirement altogether by employing a filtering approach and exploiting lightweight representations of protein structures. We review these methods next.

5.1.2 Filtering-Based Whole Structure-Based Methods

The goal of filtering methods is to rapidly eliminate structures that are not likely to share structural features with a query structure, thus making it practical to employ alignment-based methods on remaining structures. How to represent a protein structure is key to the performance of filtering-based methods, as the choice of representation directly dictates what distance metrics can then be employed to accurately and efficiently score the similarity between two structures [13, 45, 51, 52, 138, 174, 216, 235, 245, 391].

For instance, work in [298], inspired by Vassiliev knot invariants, describes the topology of a protein structure via 30 real-valued features. The scaled Guass metric (SGM) is then employed to compare two such 30-element vectors corresponding to two protein structures. A simple classification procedure employing this metric is shown to correctly identify the fold for 95 % of the structures in CATH2.4 [298]. Another method, PRIDE, compares protein structures via a fast algorithm based on the distribution of inter-atomic distances [53]. PRIDE has been shown to accurately assign query proteins to their CATH superfamily 98.8 % of the time. PRIDE and other structure analysis methods are available for the community via a web server [360]. Recently, a group of methods have exploited the fragbag representation, which is a bag-of-word (BOW) representation of a protein structure.

Utilizing techniques from machine learning and data mining, fragment libraries have been utilized to form a dictionary of "words." Utilizing libraries created

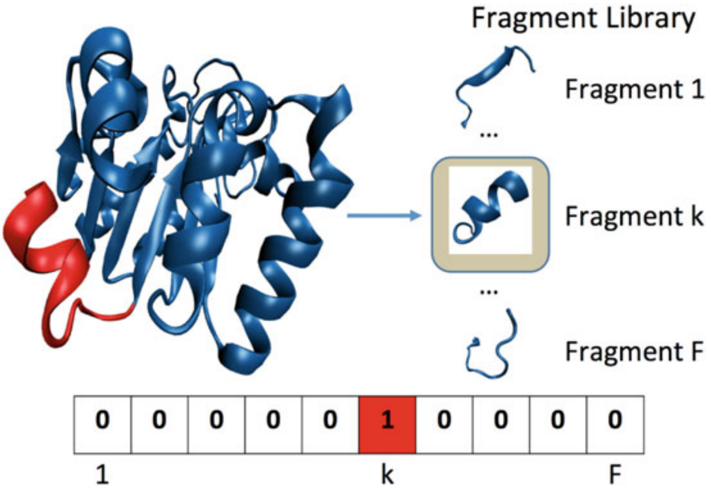


Fig. 2 Fragbag creation process. A fragment library is shown *top the right* (containing F fragments). The protein structure on the *left* (rendered with VMD [145]) is scanned one fragment at a time (first fragment highlighted in *red*) using a sliding window the same size as the fragments in the library. The library fragment that approximates the current fragment in the protein is located and a term frequency vector is constructed, with each position representing the number of times that library fragment was used in describing the protein. Fragbags are described in [45] and this figure is taken from [245]

in [181], a protein is represented as a term frequency count of the number of times each fragment is used to approximate a segment of the protein's backbone. Figure 2 from work in [245] illustrates the fragbag creation process. Fragbags when combined with a cosine similarity metric were shown to allow the fast detection of remote homologs in [45] and to provide some insight into the relationship between sequence, structure, and function [165, 269].

Work in [245] uses fragbags to construct low-dimensional categorization of the protein structure space and utilize these representations to automatically assign protein function by identifying the SCOP superfamily to which a query protein resides. This categorization is constructed using the Latent Dirichlet Allocation (LDA) model [32] made popular in topic modeling, which is an unsupervised learning approach allows further reducing the fragbag representation to a representation of 10 topics. The topic-based representation introduced in [245] is employed to identify remote homologs of a query protein structure as well as map the query to its SCOP superfamily. The latter is done via SVM, which allows reaching a classification accuracy of over 80 %.

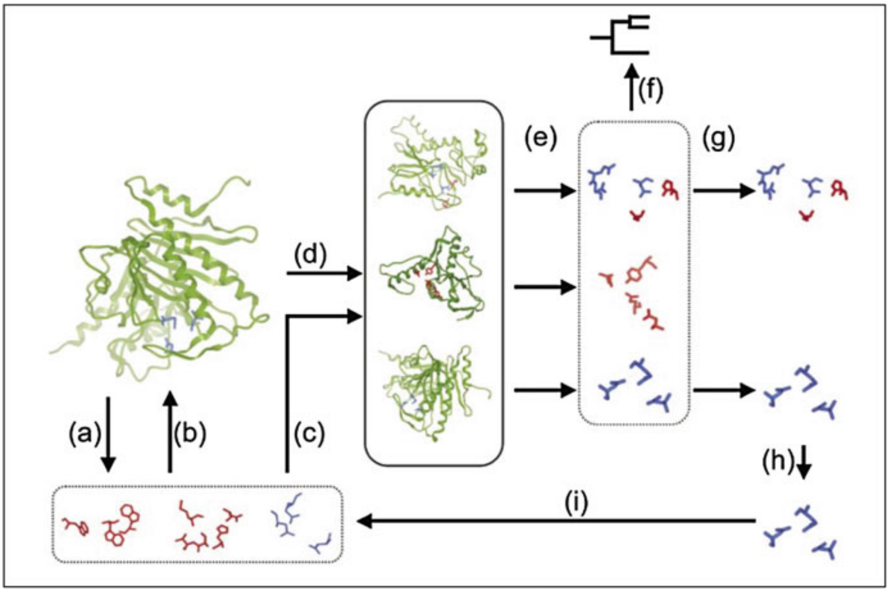


Fig. 3 Methods to detect common local substructures as shown in [254]. Template-based methods first (**a**) extract known functional motifs from an experimental structure and then (**b**) use this information to search for other structures containing these motifs. (**c**) Illustrates that expert-generated templates can also be used for search queries. The processes for pairwise methods are shown in (**d**)–(**i**). (**d**) Represents a database of protein structures which are processed to reveal local structures shown in (**e**). These structures are then clustered in (**f**) to reveal highly populated clusters extracted in (**g**) and undergo statistical analysis (**h**). The identified motifs can then be fed to (**i**) template-based methods

5.2 Substructure-Based Methods

Similar to sequence motifs, structural motifs are used to identify common structural components amongst sets of functionally related proteins. The main premise of these methods is to tabulate motifs that act as structure-function signatures and then search a given protein structure for the presence of function-known motifs.

Substructure-based methods can be categorized into two types: template-based methods and pairwise methods. The main difference between these methods is that the motifs may be pre-compiled from experimental structures or identified as shared substructures of two structures under comparison. These two groups of methods are summarized in Fig. 3. Work in [254] summarizes substructure-based methods. Many databases are now available that store structural motifs, including the Database of Structural Motifs in Proteins (DSMP) [121] and the Structural Motifs of Superfamilies (SMoS) [54].

5.3 Structure Feature-Based Supervised Learning Methods

While some of the methods already discussed use supervised learning methods, such as kNN to transfer functional annotations from neighbors to the query, others use SVMs, decision trees, and NNs. For instance, work in [139] introduces SVM-I, a method that uses structural features in lieu of sequence features to detect remote homologs, obtaining an almost four-fold improvement in runtime and improved accuracy over other SVM-based methods that utilize only protein sequence. Machine learning methods have also been applied to predict functionally important sites within a protein. In [381], Yahalom et al. train an SVM using structural features to identify catalytic residues. They utilize that these residues have specific spatial proximities and are deeply embedded in the structure. In this work they investigate other classification methods, but find that SVMs provide the best performance. Many feature-based methods that employ structural features also integrate other sources of data and are the subject of our review in Sect. 8.

5.4 State of Structure-Based Function Prediction

As the number of structures and number of functional annotations increase in structural databases, function prediction methods that utilize structure will gain even more discriminative power over methods that employ only sequence. Work in [102] provides some early evidence to this effect, by showing that structure prediction, coupled with biochemically relevant structural motifs, can outperform sequence-based methods and provide more detailed, robust function annotation of genome sequences. Computational efficiency concerns will remain a challenge. However, techniques that have originated in the data mining community are showing good potential at addressing both computational efficiency and accuracy concerns via smart representations and representation-aware distance metrics.

6 Interactions-Based Methods for Function Prediction

Direct binding can be tested at a high throughput scale in the wet laboratory via the yeast two-hybrid (screening) system (Y2H) or affinity purification coupled to mass spectrometry. These two technologies have allowed researchers to amass hundreds of thousands of protein–protein interaction (PPI) data [202]. This data can be found in databases, such as the the Database of Interacting Proteins (DIP) [379], the Biomolecular Interaction Network Database (BIND) [15], the Biological General Repository for Interaction Datasets (BioGRID) [55], the Human Protein Reference Database (HPRD) [169], the IntAct Molecular Interaction Database and the Molecular Interactions Database (MINT) [265], the MIPS Protein Interaction Resource on

Yeast (MIPS-MPact) [119], and the MIPS Mammalian Protein-Protein Interaction Database (MIPS-MPPI) [273]. These databases can be integrated, as in the Agile Protein Interaction DataAnalyzer (APID) [284], the Microbial Protein Interaction Database (MPIDB) [114], and the Protein Interaction Network Analysis (PINA) platform [69].

Other databases combine PPI data obtained via Y2H and/or affinity purification with data predicted in silico. Examples include the Michigan Molecular Interactions (MiMI) [339], Human Protein-Protein Interaction Prediction Database (PIPs) [238], Online Predicted Human Interaction Database (OPHID) [40], the online database of comprehensive Human Annotated and Predicted Protein Interactions (HAPPI) [59], Known and Predicted Protein-Protein Interactions (STRING) [334], and the Unified Human Interactome (UniHI) [161].

PPI data presents a unique opportunity for function prediction methods, particularly when the data is encoded in protein–protein interaction (PPI) networks, where vertices represent proteins and edges represent direct binding. Interactions can be used to infer functional relationships. This principle is known as "guilt by association" (GBA) [263]. Graph-theoretic concepts and algorithms can be employed on PPI networks to predict the function of a query protein. Methods that do so can be organized in four main categories:

- **Neighborhood-based methods**: these methods exploit the most dominant annotations among neighbors of a query protein in a PPI network.
- **Module-assisted methods**: these methods exploit the local topological structure of a PPI network to identify functional modules from which to infer functional annotations of unannotated proteins. Two subgroups of methods can be found in this category, clustering-based methods, which seek dense subgraphs in a PPI network, and non-clustering methods, which seek dense subgraphs via graph-theoretic concepts.
- **Global optimization-based methods**: these methods go beyond neighborhood information and consider the structure of the entire network. Optimization of an objective function allows exploiting annotations of other proteins indirectly connected to the query.
- **Association-based methods**: these methods take a complementary approach to detection of dense subgraphs, employing association rule mining to detect frequently-occurring sets of interactions.

Methods that are not module-assisted are also referred to as direct methods, as they essentially propagate functional information through the network. A detailed review of these methods for function prediction is provided in [315]. In the following, provide an up-to-date summary of these methods.

6.1 Neighborhood-Based Methods

Neighborhood-based methods transfer to a query protein the most dominant annotations among neighbors of the query in a PPI [65, 109, 137, 208, 247, 283, 312, 363]. Work in [312] was one of the earliest to demonstrate that 63 % of the interacting proteins in a yeast PPI of 2709 interactions had a common functional assignment, and 76 % were found in the same subcellular compartment. These simple statistics laid the foundation for exploiting the connectivity information in a PPI network to infer function. Functional assignment based on the majority (voting) annotation shared by direct neighbors was shown to be a viable approach [312]. This approach, deemed Majority Voting in [312], later came to be known as the nearest-neighbor voting, or basic GBA (BGBA), and is persistently shown to perform well [283] in the Critical Assessment of Function Annotation (CAFA) challenge. An earlier precursor of nearest-neighbor voting or neighbor counting methods is the chi-square method [135], which considers neighbors indirectly and assigns to a protein k functions with the k largest chi-square scores; the chi-square score for a function j and a protein P_i is defined as $S_i(j) = \frac{|n_i(j)-e_i(j)|^2}{e_i(j)}$, where $n_i(j)$ is the number of direct neighbors of protein P_i that have function j and $e_i(j) = n_i(j)xp_j$ is the expected number of neighbors with function j, with p_j denoting the fraction of proteins having function j among all proteins in the PPI network. Work in [75] shows that BGBA-based methods perform comparably to the chi-square method.

Work in [65] extends BGBA by including indirect, level-2 neighbors. The FSWeighted algorithm is proposed. The local topology of the query is compared with that of its direct and indirect neighbors to estimate functional similarity between a query and its neighbors. The experimental reliability of interactions is combined with the functional similarity to associate a weight with each neighbors. The query is then assigned the various GO terms of its neighbors, scoring each term by its weighted frequency among the neighbors. Leave-one-out cross validation shows that the performance of this method is comparable with other neighbor- and similarity-based methods, as well as Markov random field-based global optimization methods summarized below.

There are several issues with neighbor-based methods that are addressed at various levels in existing literature. For instance, the query needs to have a sufficient number of annotated neighbors in the network for a reliable prediction to be made. While early methods did not consider distances between the query and its neighbors, later methods do so, as summarized above. While GBA focuses on annotated neighbors, there is information to be exploited in interacting unannotated neighbors. Recent methods, such as those proposed in [247, 363], have begun to exploit unannotated interacting pairs in PPI networks.

Another issue concerns which neighbors to consider, and whether to restrict the GBA approach to direct neighbors, neighbors within a radius, or extend it to indirect neighbors. Work in [109] shows that indirect connections improve gene function prediction and proposes a new method based on the concept of extended GBA, where networks are extended by self-multiplication. The multiplication allows estimating the number of paths of a certain length connecting a given pair of nodes in the network, and the method in [109] weights the paths in an extended network before conducting GBA as on the original network. This network extension approach sits at the boundary of neighbor-based and global optimization methods. The approach is applied not only to PPI networks but to co-expression networks, as well, showcasing its generality for processing biological data encoded in graphs.

Yet another issue concerns the existence of mutual dependencies among neighbors of the query. If two or more neighbors have similar function, their contribution is likely to accumulate in existing neighbor-based methods. This should not be the case. Instead, dissimilar neighbors should be more important for annotation of a query. Work [137] proposes a way to take into account correlations among neighbors. In particular, the Choquet-Integral for fuzzy theory is employed to aggregate functional correlations among neighbors. The functional aggregation measures the impact of each relevant function on the final prediction and reduces the impact of repeated functional information on the prediction. The functional aggregation is employed in a new protein similarity and a new iterative prediction algorithm proposed in [137]. Evaluation of this approach shows that removing neighbor correlations results in improved performance over neighboring methods based on majority voting and sophisticated distance metrics such as the functional similarity metric proposed in [65].

Finally, the majority of neighbor-based methods ignore the scale-free property found in many biological networks, including PPI networks [3, 21]. In [209], neighbor sharing is assumed to be constrained by preferential attachment, and the Preferential Attachment based common Neighbor Distribution (PAND) method is proposed to calculate the probability of a neighbor-sharing event between any two nodes in a network. This probability distribution was shown to match very well the observed probability in simulations of scale-free networks. PAND was applied to a PPI network in [209] and shown to reveal smaller probabilities correlating with closer functional linkages between proteins. PAND-derived linkages were used to construct new networks with more functionally reliable links than links in PPI networks. Simple annotation schemes on the new networks were found to be more accurate [209].

6.2 Module-Assisted Methods

The local topological structures and properties of PPI networks are subject to theoretical investigation and empirical exploration via ideas from network science. Module-assisted methods seek to identify local topological structures that can represent functional modules in a PPI network. Clustering-based methods rely on clustering to identify dense regions with a large number of connections in PPI networks, whereas non-clustering methods employ graph-theoretic concepts and algorithms to identify local topological structures. We review each next.

6.2.1 Clustering-Based Methods

Clustering methods focus on finding dense regions with a large number of connections as a way of identifying functional modules representing protein–protein complexes/assemblies [23, 330]. MCODE [23] uses a vertex weighting scheme based on the clustering coefficient to measure the cliquishness of the neighborhood of a node. Work in [330] proposes two clustering algorithms. The super paramagnetic clustering algorithm [30] is a physics-inspired hierarchical clustering algorithm. The Monte Carlo algorithm instead maximizes the density of predicted clusters. The Markov clustering (MCL) algorithm is proposed in [94, 280], whereas the highly connected subgraph (HCS) algorithm is proposed in [128]. HCS is a graph-theoretic algorithm that separates a graph into several subgraphs using minimum cuts. A cost-based local search based on tabu search metaheuristic is proposed in [171].

Other methods employ classic clustering algorithms after defining a similarity measure that takes into account the interactions of a protein in the network [11, 42, 113, 228, 302]. The SL method in [302] employs the number of common neighbors to define the similarity between two proteins and then uses k-means to partition the nodes in a PPI network into different groups/assemblies. Following work in [228] modifies the similarity measure via a weighted form of the mutual clustering coefficient approach [113]. Work in [11, 42] uses hierarchical clustering. In [11], the shortest path distance between two proteins is used to estimate their similarity, while work in [42] uses the Czekanovski-Dice (CD) metric.

The CD distance between two proteins u and v is based on the number of neighbors they share and is measured as: $\frac{|N_u \delta N_v|}{|N_u \cup N_v| + |N_u \cap N_v|}$, where N_* refers to the direct neighbors of a vertex, and δ refers to the symmetric difference between two sets. Work in [42] proposes the PRODISTIN method, which employs the BIONJ algorithm (an improved version of the popular neighbor-joining algorithm) [106] to cluster proteins in a PPI based on their CD distance. The BIONJ algorithm produces a hierarchical classification tree. A PRODISTIN functional component is the largest subtree that contains at least three proteins with the same function and has at least 50 % of its annotated members sharing that function. This function is transferred to the unannotated proteins in the functional component. It is worth noting that the neighbor-based method in [65] proposes a new CD-based functional similarity that penalizes the similarity weights between protein pairs when any of them have few direct neighbors.

Work in [364] shows that clustering-based methods can produce many false positives. To remedy this issue, the EVDENSE method is proposed in [364] to efficiently mine frequent dense subgraphs in a PPI network. EVDENSE produces frequent dense patterns by extending vertices and by using relative support. Improved performance is reported over other clustering-based methods.

A detailed analysis of clustering-based methods is conducted in [325] to evaluate the hypothesis that dense clusters correspond to functional modules. Six different clustering algorithms are applied to a yeast PPI network, and evaluation shows that the performance of these algorithms is dependent on the topological characteristics of the network. For the specific task of function prediction, a non-clustering, BGBA approach outperforms the clustering algorithms. Guidelines are provided in [325] to evaluate and justify novel clustering methods for biological networks.

One issue of clustering-based methods regarding the dynamic process in clustering is addressed in the PClustering method proposed in [301]. Saini and Hou track function appearance across all relevant clusters rather than just the cluster to where the query is mapped by a specific clustering algorithm. A recursive clustering algorithm reclusters until a cluster is obtained where the query gets separated from the rest of the other proteins. The recursion tree tracks how the clusters are split down to the leaf where the query is separated from all other proteins. The particular path from the root to this leaf is then inspected to accumulate all functions of all proteins in the path. This set is the list of relevant functions that can be assigned to the query. The prediction for the query is then made based on how the proteins in the path, with relevant functions, are split during the recursive clustering process. A score is proposed to select functions that are stable in terms of their frequency across the clusters in the path. PClustering is compared to PRODISTIN [42], MCL [94], SL [302], Chi-square [135], Majority Voting [312], and the FSWeighted method [65] on the yeast PPI dataset and shown to outperform many of these methods, particularly on being able to accurately predict functions of more unannotated proteins.

6.2.2 Non-clustering, Graph-Theoretic Methods

Clustering-based methods essentially aim to uncover communities in a network. A different group of module assisted methods circumvent clustering and instead employ concepts and algorithms from network science for community detection. A community is a more general concept than a cluster, and various methods exist on community detection in networks, a review of which is beyond the scope of this paper. Here we summarize recent methods that uncover functional modules by exploiting the concept of communities and then employ such modules for function annotation.

Work in [222] introduces the concept of k-partite "protein" cliques as functionally coherent but not necessarily dense subgraphs. This concept is more suitable for PPI networks, which are known to be non-uniform in subgraph density for reasons that are often artifacts of wet-laboratory studies. Briefly, a k-partite protein clique is a maximal k-partite clique comprising two or more non-overlapping subsets between any two of which full interactions are exhibited. In [222], a PPI network is transformed into induced k-partite graphs, where edges exist only between the partites. A maximal k-partite clique mining (MaCMik) algorithm is proposed to enumerate maximal k -partite cliques on these k-partite graphs. MaCMik is applied to a yeast PPI network, and unusually high functional coherence is observed in the k-partite cliques. This direction of work suggests that graph-theoretic concepts can be more powerful at capturing the concept of functional modules and in turn assist with function prediction.

While work in [222] restricts itself to the proposal of a new concept for a functional module, work in [199] shows that the concept of a community can be exploited for function prediction. In general, a cost function is designed to measure the extent to which a subgraph constitutes a community, and then optimization algorithms partition the graph into subgraphs that optimize the cost function. The modularity (Q) measure is one example of a popular cost function that measures the relative density of intra-community connectivity compared to a randomly re-wired counterpart with the same degree of nodes. The conformational space annealing (CSA) algorithm [200, 201] is employed in [222] to detect maximum-Q subgraphs in the yeast PPI network. Figure 4 showcases the ability of the CSA algorithm to detect more subgraphs than a baseline, popular simulated annealing approach.

After the high-modularity subgraphs are detected, Random Forest (RF) is then employed, representing each protein as a vector of features generated only from the network community (including which communities the neighbors belong to and their functions). The resulting RF-comm-CSA method is compared to MRF-based methods in [77, 163], neighborhood enrichment methods [315, 325], and the Majority Voting method [312]. RF-comm-CSA is reported to achieve the best performance, followed by the MRF-based method in [163].

6.3 Global Optimization-Based Methods

To overcome these setbacks, global optimization methods consider the full topology of the network and employ techniques, such as Markov random fields, simulated annealing, and network flow [57, 77, 204, 252, 355]. The Markov random field (MRF) method proposed in [77] computes the probability that a protein has a function given the functions of all other proteins in the interaction dataset. MRFs are particularly suitable for modeling the probability that a query has a certain function by capturing the local dependency of the query on its neighbors in a PPI network. The latter is in essence the GBA principle. The Markov property is valid here, as the function of the query is assumed to be independent of all the other proteins given its neighbors in the PPI network. The MRF method in [77] is shown to be more sensitive at a given specificity than neighbor-based and chi-square methods.

Work in [355] proposes a simulated annealing method, which was later shown in [76] to be a special case of the MRF method [77]. The approach proposed in [204] is identical to the MRF method in [77]. More recent work in [57] proposes a bagging MRF-based method (BMRF). The method follows a maximum a posteriori principle to form a novel network score that considers interactions in a PPI. The score is used by the method to search for subnetworks with maximal scores. A bagging scheme based on bootstrapping samples is also employed to statistically select high-confidence subnetworks. While work in [57] applies BMRF to identify subnetworks associated with breast cancer progression, later work in [317] shares the BMRF-Net software with the community to identify subnetworks of interest in a PPI by the BMRF method.

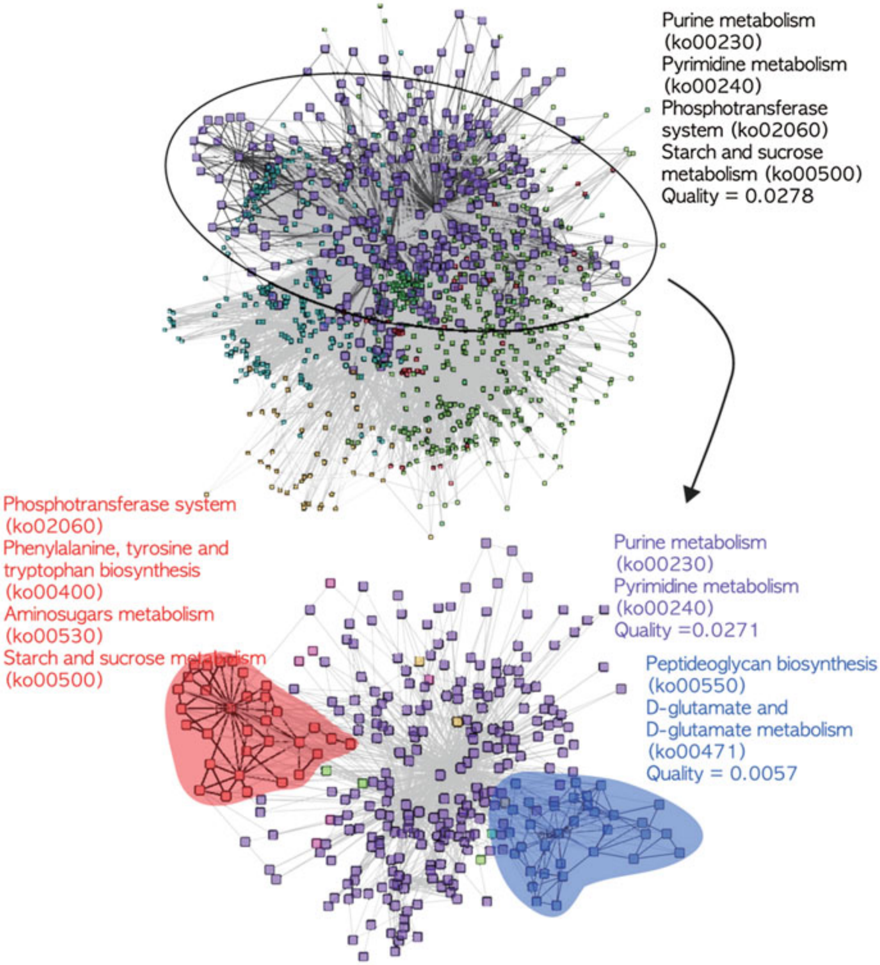

Purine metabolism (ko00230)
Pyrimidine metabolism (ko00240)
Phosphotransferase system (ko02060)
Starch and sucrose metabolism (ko00500)
Quality = 0.0278
Phosphotransferase system (ko02060)
Phenylalanine, tyrosine and tryptophan biosynthesis (ko00400)
Aminosugars metabolism (ko00530)
Starch and sucrose metabolism (ko00500)
Purine metabolism (ko00230)
Pyrimidine metabolism (ko00240)
Quality =0.0271
Peptideoglycan biosynthesis (ko00550)
D-glutamate and D-glutamate metabolism (ko00471)
Quality = 0.0057

Fig. 4 In [199], nodes of one community obtained by simulated annealing are split into three communities by the CSA algorithm. Two of these are in *red and blue shaded areas*, and the third is drawn with *large squares*. Within each of these communities, meaningful functional clusters of KEGG pathway annotations with P-value less than 10^{-4} are listed

While the MRF-based method in [77] does not consider unannotated proteins when training the regression model for parameter estimation, work in [187] uses the same MRF model but then applies adaptive Markov chain MC (MCMC) to draw samples from the joint posterior. This approach gives an AUC value of 0.8915 on 90 GO terms compared to the 0.7578 reported in [77] and 0.7867 reported in [204]. Work in [76] extends the MRF model by integrating various data sources, such as PPI data, expression profiles, protein complex data, and domain information.

Graphical models such as MRFs continue to solicit interest for predicting protein function from biological networks such as PPI networks. A recent review in [338] focuses on MRFs and conditional random fields (CRFs) and their applications for predicting protein function and protein structure. Interestingly, the review finds that, while CRFs have become popular in protein structure prediction and protein design, they have yet to solicit interest in the function prediction community.

Work in [163, 252] represents a complementary approach based on the concept of flow. Karaoz and co-workers visualize propagation diagrams to illustrate the flow of functional evidence from annotated to unannotated proteins in a PPI network [163]. However, it is Nabieva and co-workers that operationalize on the concept of flow in the Functional Flow method [252]. The method simulates functional flow between proteins. A protein annotated with a specific function is assigned an infinite potential for that function, whereas an unannotated protein is assigned a 0 potential. Functions then flow from proteins with higher potentials to their direct neighbors with lower potential. The amount of flow depends on the reliability of the interactions.

6.4 Association-Based and Other Mining Methods

Association-based methods employ association rule mining to detect frequently occurring sets of interactions. The local topology of a query in a PPI, whether restricted to direct and/or indirect neighbors, or extended to the more general concept of a community, can be encoded via features in a feature-based representation of the query and other proteins in the PPI network. Such representations open the way for application of popular supervised learning algorithms for function prediction. For instance, work in [260] employs logistic regression on representations that use only functional annotations of the direct neighbors of a protein. Work in [64] mines a PPI for frequent functional association patterns. The set of functions that an annotated protein performs is assigned to the protein as a label, and a functional association pattern is represented as a labeled subgraph. A frequent labeled subgraph mining algorithm efficiently searches for functional association patterns in a PPI network. The algorithm increases the size of frequent patterns one node at a time by selective joining, while simplifying the network by a priori pruning. The algorithm is reported to identify more than 1400 frequent functional association patterns in the yeast PPI network. Function prediction is carried out by matching the subgraph that contains the query with the frequent patterns analogous to it. Leave-one-out cross validation shows that this function prediction approach outperforms neighbor-based methods [64].

In general, however, these methods do not restrict themselves to features extracted only from the local topology of a node in a PPI network, but integrate various data sources to represent proteins. For instance, work in [50] proposes three probabilistic scores, MIS, SEQ, and NET, to combine protein sequence, function association, protein–protein interactions, and gene–gene interaction networks. The MIS score is generated from homologs found for a query via PSI-BLAST and association rule between GO terms learned by mining the SWISS-PROT database. The SEQ score is based on sequences, and the NET score is generated from PPI and gene–gene interaction networks. The three scores are combined in the Statistical Multiple Integrative Scoring System (SMISS), which is reported to outperform three baseline methods that combine profile-sequence homology search, profile–profile homology search, and domain co-occurrence network [366] on the CAFA1 dataset.

A detailed description of data integration methods is provided in Sect. 8.

7 Gene Expression-Based Methods for Function Prediction

A complementary source of data that can be used for function prediction are gene expression data. cDNA microarray technology allows measuring the amount of protein a gene makes at a given time under specific conditions. cDNA chains can be designed to bind complementary mRNA so as to detect the transcription of specific genes. Gene expression data, also referred to as microarray data, can be measured in time, as well. Often, expression data measured across different labs and under different conditions can vary, and standard normalization strategies help eliminate discrepancies [58] and allow the employment of machine learning techniques to mine such data. Gene expression data are typically viewed in matrix format, with genes in rows and gene expressions under different conditions in columns. By now, many such data exist for different organisms. The Stanford MicroArray Database [316], now retired, contained expression data for genes in the human genome. Similar data can now be found in databases, such as the NCBI Gene Expression Omnibus [22, 84] or EMBL-EBI's ArrayExpress [178].

Microarray data are prime for machine learning methods to detect changes in gene expressions indicating the presence of specific types of diseases. Such data can also be used for function prediction, as similarities between expression profiles of genes can indicate functional similarities. This hypothesis was investigated early on in [362], where 40,000 genes were examined for co-expression with five genes known to be associated with prostate cancer. The guilt-by-association principle was employed to identify uncharacterized genes significantly co-expressed with at least one known prostate gene. This approach detected eight novel genes associated with prostate cancer. While technically this study pursued the exploitation of gene expression data to detect gene signatures of disease, it opened the way to more sophisticated machine learning methods for the more detailed problem of function prediction. Such methods can be organized in the following three categories:

- **Unsupervised learning methods**: these methods cluster expression profiles to identify genes with similar profiles that can be hypothesized to share functional annotations.

(continued)

- **Supervised learning methods**: Predicting function from gene expression data is a natural learning problem in the supervised setting, and these methods investigate classifiers, such as SVMs, naive Bayes, NNs, and more.
- **Temporal analysis methods**: these methods exploit the ability to measure gene expression in time during, for instance, a disease. Temporal expression data can also be used by classifiers to predict the function of unannotated genes.

7.1 Unsupervised Learning of Gene Expression Data for Function Prediction

Many different clustering algorithms can be used to organize gene expression profiles, including clustering algorithms specifically designed for gene expression data, such as CAST [25]. The issue, however, is what to do with the clusters and which cluster to employ for function annotation. Measures such as majority [122] have not been effective [395]. Work in [85] demonstrated an effective approach relying on hierarchical average-linkage clustering with a variation of the correlation coefficient as a similarity measure. Analysis showed that genes mapped to the same cluster were for the most part involved in common cellular processes, directly validating the hypothesis that clusters of co-expressed genes are also functionally coherent. This result is often credited as bringing clustering to the forefront of techniques for analysis of biological data.

The results of the relevance of clustering for functional annotation motivated the development of a novel clustering algorithm, the Cluster Affinity Search Technique (CAST) [25]. CAST proceeds in two phases, an add phase where elements with high affinity to the current cluster are added to the cluster, and a remove phase, where elements with low affinity are removed from the cluster. Clusters are constructed one by one, and the algorithm terminates when no changes occur. Evaluation on static and temporal gene expression data showed that CAST is able to preserve functional categories. The ability of the algorithm to extract knowledge about diseases from expression data has also been demonstrated, and has been employed in other works to study gene signatures of disease [24, 333].

Work in [259] focused on removing noisy and redundant dimensions from expression profiles via latent semantic indexing [205]. The retained dimensions are then employed for clustering based on the concept of neighborhood, where intra-cluster similarity needs to be significantly higher than inter-cluster similarity. Unannotated genes in a cluster then are assigned the majority annotation of the characterized genes in the cluster.

The known issues with clustering regarding how to define similarity and how many clusters to extract have prompted many researchers to pursue additional directions. For instance, work in [395] increases confidence in the results via the novel ontology-based pattern identification (OPI) strategy. Briefly, all decisions such as attribute weights, choice of similarity threshold, choice of mean or median, and other decisions that need to be made in clustering algorithms are embedded in a Euclidean space. Then, a hill climbing algorithm is employed to navigate this space in order to find decisions that are optimal with regard to an objective function encoding expected characteristics of optimal clustering. The hill climbing algorithm minimizes this function for all the GO functional categories and in the process identifies the best cluster for each category. As in earlier work, unannotated genes in a cluster are functionally linked to annotated genes in the same cluster. OPI has been shown able to identify more statistically-significant clusters than other related work employing the k-means clustering algorithm [297]. OPI has been additionally validated. About 12 of the 50 genes predicted by OPI to have the antigenic variation have now been verified.

Dealing with cluster overlaps is another way to increase confidence in the clustering results. Work in [378] employs different clustering algorithms and annotates a cluster with the functional class of the least p-value, as calculated from the fractions of the different functional classes in a cluster. An unannotated gene is then assigned the functional class of the cluster to which it is mapped, and the assignment is also associated a confidence value based on the p-value of the cluster. Consensus clustering is proposed in [246] as an analysis approach to assist clustering algorithms. Consensus clustering, together with resampling, is proposed to represent the consensus over multiple runs of clustering algorithm with random restarts, such as K-means, model-based Bayesian clustering, and self-organizing maps (SOM), which are typically sensitive to initial conditions.

The method in [333] fuses consensus clustering with ideas proposed in [378] to carry out both robust and consensus clustering of gene expression data, as well as assign statistical significance to clusters from known gene functions. While the method in [246] perturbs gene expression data for a single algorithm, the method in [333] uses different clustering algorithms. This method proposes a robust clustering algorithm, which seeks maximum agreement across different clustering methods by reporting only the co-clustered genes grouped together by all the different algorithms. To address the issue of robust clustering discarding gene expression vectors if only one clustering method performs badly, consensus clustering is proposed, which seeks a minimum agreement. An objective function is defined to reward clusters with instances of high agreement and penalize clusters with instances of low agreement. The function is minimized via simulated annealing. Consensus clustering is reported to improve upon the performance of individual clustering methods, as measured by the weighted-k measure [7]. Clusters identified for ten functional classes in [333] were also more likely to be annotated with the same classes by consensus clustering than individual clustering methods.

The functional coherence of gene expression clusters is called into question in [394], which investigates a graph-theoretic approach. Genes are encoded as vertices, and edges connect genes with correlated expression profiles. Shortest paths in the graph allow identifying transitively related genes. A simple experiment is conducted, where shortest paths between genes of the same GO category are analyzed to check if the genes in these paths are annotated with the same GO category or a parent or child function in the *S. cerevisiae* Genome Database (SGD) [82]. The analysis shows that high accuracy is obtained for mitochondrial and cytoplasmic genes, but medium accuracy is obtained for nuclear genes. The graph-theoretic approach in [394] additionally provided functional annotations to 146 genes that were weakly correlated to other genes.

Another approach is employed in [229] to relax the functional coherence requirement of a cluster. The latter may be weak if one considers all conditions, as some may act as noise and strong correlation can be obtained upon removal of such entries. Biclustering or coclustering is employed to address this. Briefly, biclustering is a specific type of sub-space clustering and refers to the simultaneous clustering of both rows and columns of a data matrix [127]. Two-way analysis of variance is then used to identify constant valued sub-matrices. While many methods pursued biclustering for gene expression data [62, 384], it was work in [43] that demonstrated a simulated annealing approach to biclustering to perform well on yeast cell cycle data sets. The two largest clusters were found to contain largely members of two different families, the ribosomal proteins and the nucleotide metabolism proteins. The method in [221] further demonstrated the potential of biclustering for functional annotations. The structure of the GO hierarchy is incorporated in the hierarchical biclustering process. Genes are first clustered via hierarchical clustering, and then each node in the hierarchy is annotated with the GO functional class with which it is most enriched.

While much progress has been made in clustering algorithms, many issues inherent to clustering remain. One issue, in particular, pertains to the inability of clustering algorithms to exploit already labeled instances. Supervised learning methods exploit labeled data, and we review such methods for gene expression data next.

7.2 Supervised Learning of Gene Expression Data for Function Prediction

Gene expression data are subjected to a classification setting in [41], where three different classifiers are compared to learn functions from yeast gene expression data: Parzen's window, Fisher's linear discriminant analysis, two decision tree classifiers (C4.5 and MOC), and SVMs with different kernels. Comparative analysis demonstrates SVM with the radial basis kernel to perform best. Work in [192] compares SVM to kNN classifiers on gene expression data. The cosine similarity measure is employed, as it better captures the shape of an expression vector in high-dimensional space rather than its magnitude. kNN is reported to be better at predicting the *m* most appropriate classes for a test gene over SVM.

Multilayer perceptrons are employed in [237] to learn 96 different functional classes from annotated/labeled yeast gene expression data. Better performance is obtained over work in [41], but three sources of errors are identified: class size, class heterogeneity, and Borges effect (termed in [237] to indicate simultaneous membership of a gene in different functional classes). An iterative learning procedure is proposed in [237] to address these three sources of error better than a one-pass learning procedure.

Work in [258] employs multiple expression data sets for learning with SVMs and presented a strategy to select the most informative data set for learning individual classes. The learning cost savings measure introduced in [41] is used to show that blindly combining different data sets is not optimal. A hill-climbing algorithm is proposed instead to incrementally add the data set that provides the maximum learning cost savings until a maximum is reached. Comparison with other classification methods showed this approach to be superior.

A different direction is investigated in [388], namely, that of training on larger mammalian expression data sets (the earlier works above focused on *S. cerevisiae* or *Caenorhabditis elegans*). An SVM is employed to learn each of the 992 GO biological process categories and classify 10,000 unannotated genes. Performance is reported to be mixed, suggesting that more sophisticated techniques may be needed for mammalian gene expression data sets. Specifically, while SVMs are at the moment the state of the art in classification of gene expression data [250], there is room for investigating different classification techniques, such as boosting, active learning [81], and more.

7.3 Temporal Gene Expression Data Analysis for Function Prediction

Temporal gene expression data provide dynamic information on the simultaneous expression levels of genes, effectively providing a dynamic picture of what goes on in the cell via expression measurements. Both unsupervised and supervised learning techniques have been applied to temporal expression data. Work in [19, 46, 96, 132, 155, 243] exposes challenges posed by time-series expression data to clustering algorithms, such as similarity measures, co-clustering, short profile lengths, and unevenly sampled genes. The issue of possible time offsets in the expression of different but functionally associated genes is another one that still challenges temporal gene expression analysis. Some progress has been made.

Temporal gene expression data are used in [148, 193] to learn GO biological process annotations for unannotated genes. The gene expression profiles are transformed into attribute-valued vectors. Attributes constructed by calculating the increase or decrease of expression values between two instances separated by an interval of three time points. Three possible values are assigned, high, medium, and low. This mapping is central to the ability to use set-based classifiers, which can only robustly handle nominal attributes. The classifier in [148] achieves a cross-validation AUC score of 0.8. Testing on the human serum response expression data set results in labels being correctly predicted for 211 of the 213 genes.

Inductive logic programming and description logics are proposed in [14] for learning the classification rule set from temporal gene expression patterns. Work in [241] presents another rule-based classification model which outperforms the approach in [148]. HMMs are employed in [78] to model the interdependence between conditions and the dependence of the functional class on them. A dual HMM modeling both expression values and experiment order is shown to perform best on yeast gene expression data. Another statistical approach, Mixture Functional Discriminant Analysis (MFDA), is proposed in [118] to operationalize upon the observation that temporal profiles of genes belonging to the same functional class are highly similar. Each individual class is modeled as a mixture of sub-classes, and the Expectation Maximizaton (EM) algorithm [81] is used to learn the parameters of the model. MFDA is marginally better over other discriminant analysis methods on yeast cell cycle expression data.

Work in [353] injects evolution and evaluates the hypothesis that the conservation of co-expression between pairs of genes that share an evolutionary history can enable more confident prediction of their functional association and pathways in which they are involved. The evaluation is conducted on *S. cerevisiae* and *C. elegans*, using correlation as a measure of co-expression. Two types of co-expression conservation are defined: Paralogous conservation, which refers to two pairs of genes (A, B) and (A', B') in the same organism, where A is homologous to A' and B is homologous to B; Orthologous conservation, where the two pairs (A, B) and (A', B') belong to different organisms. A correlation threshold of 0.6 results in an accuracy of 93 % and 82 % on orthologous and paralogous conservation in *S. cerevisiae*.

Research on temporal gene expression analysis is very active. Work in [244] proposes a clustering algorithm capable of handling unevenly sampled temporal gene expression data. A novel dissimilarity measure is proposed in [72] to assist graph-based clustering methods on temporal gene expression data. Smoothing spline derivatives are combined with hierarchical and partitioning clustering algorithms in [74] to capture the effects of fasting on the mouse liver. Dynamic clustering is shown in [215] to statistically estimate the optimal number of clusters and distinguish significant clusters from noise. A novel sub-space clustering algorithm is proposed in [341]. A detailed review of the state of analysis methods for temporal gene expression data is presented in [20].

8 Data Integration Methods for Protein Function Prediction

Integrative methods exploit and integrate heterogeneous data to improve the accuracy of function prediction. This category of methods operates under the umbrella of machine learning and can be organized in mainly four categories:

- **Vector-space integration methods**: these methods combine features extracted from different sources of biological data into one typically long feature vector. The transformation then allows investigating function prediction under the umbrella of machine learning.
- **Classifier integration methods**: these methods do not combine features into one long vector but instead train separate classifiers on separate feature vectors extracted from the different biological data. The results of the classifiers are then combined via the ensemble approach.
- **Kernel integration methods**: these methods employ a special similarity matrix known as the kernel matrix. A kernel matrix records the pairwise similarities between the proteins under investigation. Data sources can be kept separate, with a kernel matrix for each data source. The kernel matrices can also be combined via basic algebraic operations. Standard supervised classifiers, such as SMV and kNN, can be then used.
- **Network integration methods**: these methods rely on encoding pairwise similarities as edges of a graph. Different graphs can be constructed for the different data sources under consideration. The graphs can be then unified, and function prediction can proceed via generative or discriminative machine learning models.

8.1 Vector-Space Integration Methods

Methods already described in this review that combine sequence, physico-chemical, and secondary structure information about a protein into one long feature vector fall in this category. Other methods that combined genomic context, phylogenetic profiles, and phylogenetic trees also belong to this category. Here we describe some recent methods that combine additional data sources to improve the accuracy of function prediction. However, these methods integrate data from essentially the same source. The first employ data that are extracted from the amino-acid sequence, whereas the latter employ data that are extracted from the evolutionary history of a protein. The ProtFun method proposed in [153] and then applied to predicting GO annotations in [154], described as a feature-based method for inferring function from sequence, is indeed a vector-space integration method. We recall, as described in Sect. 2.3, that ProtFun integrates sequence data with post-translational modifications, such as N- and O-glycosalization, phosphorylation, cleavage of N-terminal signal peptides, and other modifications and sorting events that a protein is subjected to before performing its function.

Work in [223] takes a unique approach to function prediction by focusing on proteins with intrinsically disordered regions (IDRs). Some studies estimate that between 30 and 60 % of eukaryotic proteins contain long stretches of IDRs, and work in [223] investigates the extent to which function can be inferred from information hidden in these regions. Specifically, pattern analysis of the distribution of IDRs in human protein sequences shows that the functions of intrinsically disordered proteins are length- and position-dependent. A total of 122 features are extracted from a protein in [223], and the features cover 14 different sources of biological information about a protein. The latter range from sequence-based features, such as sequence length, molecular weight, average hydrophobicity, charge, and more, to transmembrane-based features, such as number of transmembrane residues, percentage of N-terminal and C-terminal residues, and more, to secondary-structure features, Pest region features, phosphorylation features, O- and N- glycosylation features, and peptide features, and disorder-related features. The latter can be easily extracted from sequence via tools, such as disEMBL [149]. Correlations between the 122 features are investigated in [223], and multidimensional scaling (MDS) is applied to see how organization in a three-dimensional embedded space. Visualization of the three-dimensional embedded space obtained by MDS shows not only correlations between features extracted from the same category/source of biological information, but also correlations across features of different categories.

The attributes are valued and recorded in one long feature vector for each protein, and an SVM is trained on 26 GO categories. Improvements in accuracy are observed over a version of the classifier without the disorder features and ProtFun [153] and the method with no disorder. The individual contribution of each feature is also estimated via loss of classification accuracy upon feature removal. Significant improvements are observed for specific functional categories, such as kinases, phosphorylation, growth factors, and helicases.

Work [367] proposes the CombFunc method, which incorporates the ConFunc method proposed by the same authors (ConFunc is a homology-based method described in Sect. 2.1) and other methods that use sequence, gene expression, and protein–protein interaction data. Three categories of features are employed, sequence-based features, protein–protein interactions, and gene co-expression. Sequence-based features include those used by ConFunc, the E-values of the top annotated BLAST and PSI-BLASt hits, the sequence identity between the query and the top hits, and the sequence coverage of the query by the top hits. The i-score proposed in the GOtcha method [236] to take into account the annotations of multiple sequences returned by PSI-BLAST is included in the sequence-based features.

Other sequence-based sources of data include domain information about they query, as obtained with Interpro [146] and structures homologous to the query in the fold library of Phyre2 [167]. The domains and corresponding GO term annotations of the domains identified by InterPro are used to encode additional features. For each of the identified GO terms, the lowest E-value of a domain hit annotated with that term is recorded and added to the feature set. Pfam domain combinations [329] are also used to make predictions as in [103]. In this case, only one feature is added to the feature set, 1 if predicted by the method and 0 otherwise. Features from the Phyre2 fold library employ GO terms present in the top annotated hit and the probability score from the HHsearch [323] between the query and hit and the sequence coverage of the query by the hit.

For the GO terms identified by the interactome analysis, the features added to the feature set are the fraction of direct and indirect neighbor annotated with each term. For GO terms identified from the gene expression data, the features added to the feature set are the fraction of co-expressed genes annotation with the particular term, as well as the minimum, average, and minimum mutual rank and correlation coefficients of the co-expressed genes. In this manner, each protein sequence is transformed into a 30-dimensional feature vector.

Three different SVMs are employed in CombFunc for the different levels of the GO hierarchy under the molecular function and biological process categories. One SVM considers only terms one level below the root (for instance, catalytic activity or binding for the molecular function category). Another SVM considers the terms in the next two levels, and the third SVM considers the rest of the more specific terms. The reason for training three separate classifiers is due to the insight that potentially different subsets of features may be correlated with different levels in the GO hierarchy.

CombFunc is evaluated on predictions of GO molecular function terms on a set of 6686 proteins. UniProt-GOA annotations are extracted for the proteins, but only 5000 of them are used for training, with the rest used for testing. On the testing data set, CombFunc obtains a precision of 0.71 and recall of 0.64. Performance on prediction of GO biological process terms is slightly lower, with a precision of 0.74 and recall of 0.41.

8.2 Classifier Integration Methods

Integrating data by combining it into essentially a common representation often results in information loss [195, 397]. For this reason, classifier and kernel integration methods are pursued as better alternatives over vector space integration methods.

Integration of different classifiers has been investigated for sequence-based function prediction. We recall that the GoPred method proposed in [304] and described in Sect. 2.4 combines different classifiers and then evaluates the performance of different combination strategies, such as majority voting, mean, weighted mean, and addition.

Work in [292] demonstrates that competitive or superior performance can be obtained on prediction of top-level classes in the FunCat taxonomy [300] by using an ensemble of classifiers for data integration than by vector space integration and kernel fusion-based methods. The data sources considered in [292] are protein sequence, gene expression data, domain information, and protein–protein interactions. In [292], binary SVMs are trained on each data source, and three combination strategies are evaluated, weighted majority voting, naive Bayes, and decision templates [190]. It is worth noting that the naive Bayes and logistic regression combination strategy for integrating the outputs of several SVMs trained with different data sources and kernels has already been proposed in [117] and [261], respectively, to produce probabilistic outputs corresponding to GO terms.

Work in [117] is part of the MouseFunc function prediction project and integrates new data sources not previously considered such as disease, phenotype, and phylogenetic profiles in training of three different SVMs. Three combination strategies are evaluated, bootstrap aggregation, hierarchical Bayesian, and naive Bayes combination. One of the results in [117] is that the naive Bayes combination of the per-dataset SVMs outperforms a single SVM classifier for several GO terms. A comparison of the different combination strategies shows that the naive Bayes performs best, followed by the hierarchical Bayesian over the bootstrap aggregation.

Work in [261] proposes "reconciliation" to address the drawback of making predictions for GO terms independently; the latter often results in assigning to a query protein a set of GO terms that are inconsistent with one another; that is, that do not obey the GO hierarchy. In [261], the different, independent predictions are calibrated and combined to obtain a set of probabilistic predictions consistent with the GO topology. A total of 11 distinct reconciliation techniques are considered to combine predictions for each term obtained from different SVM classifiers with different kernels. The techniques are three heuristic ones, four variants of a Bayesian network, an extension of logistic regression to the structured case, and three novel projection techniques, such as isotonic regression and two variants of a Kullback–Leibler projection technique. Isotonic regression is shown to perform best in being able to use the constraints from the GO topology.

Work in [386] addresses the multi-label setting in GO annotation predictions. A transductive multi-label classifier (TMC) and a transductive multilabel ensemble classifier (TMEC) are proposed to predict multiple GO terms for unannotated proteins. The TMC is based on a bidirected birelational graph with edges connecting protein pairs, function pairs, and protein–function pairs. An interfunction similarity measure is used to encode function–function edges. Protein–protein similarity is specific to a data source. Directionality is added to the graph to avoid issues of annotation change and function label override. The TMC uses network propagation via a nonsymmetric propagation matrix on the resulting directed bidirectional graph by optimizing local and global consistency functions [393]. Three TMCs are trained simultaneously, each one considering a different data source. Sequence, protein–protein interaction, and gene expression data are considered. The TMEC then combines the output of the three classifiers via a weighted majority vote scheme, where a classifier's influence on determining a particular GO term for a protein is proportional to its confidence on that prediction. Evaluation is carried out on predicting biological process GO categories on benchmark yeast, human, and fly protein data sets proposed originally in [248]. Comparisons on Ranking Loss, Coverage, and AUC with related multi-label methods, such as PfunBG [156], GRF [389], SW [248], MKL-Sum [337], and MKL-SA [44], show comparable or superior performance by the TMEC [386].

While work in [117, 261, 386] ignores the hierarchy of the taxonomy and then relies on ensemble techniques to reconcile conflicting predictions, work in [307, 352] either proposes classifiers that obey the GO hierarchy directly, or ensemble techniques that make final decisions based on the hierarchy of the taxonomy. For instance, hierarchical multi-label decision trees are combined via bagging in [307]. Hierarchical multi-label decision trees are intuitive in that they exploit the ability of the decision tree model to obey intrinsic hierarchy in the target taxonomy. Essentially, the query gene can be compared via sequence similarity to all genes annotated with a specific GO term, and the tree proceeds down the GO hierarchy. Bagging is shown in [307] to best combine decision tree classifiers over random forest and boosting. In [352], the topology of the GO hierarchy is not considered in the classifiers, but it is directly integrated in a novel ensemble technique.

The key observation employed in [352] is that an annotation for a class/node in the hierarchy automatically transfers to the ancestors. This is also known as the "true path" rule (TPR), which governs hierarchical taxonomies, such as GO and FunCat. The TPR ensemble technique proposed in [352] is a hierarchical ensemble algorithm that puts together predictions made each node by local base classifiers to realize an ensemble that obeys TPR. As in ensemble methods, the classifiers are trained independently, and they make predictions for their corresponding nodes. The algorithm then combines these predictions via an information propagation mechanism that can be characterized as a two-way asymmetric information flow. The information traverses the graph-structured ensemble. While positive predictions for a node influence in a recursive way its ancestors, negative predictions influence the offsprings. This is related to work in [157], where negative information propagates from a node to its offspring. In [352], in addition, positive information propagates from a node to its ancestors.

Seven biomolecular data sources are integrated in [352], such as sequence, domain, phylogenetic, protein–protein interaction, and gene expression data. SVMs and logistic regression are used as base classifiers. Evaluation of the hierarchical ensemble technique is carried out on *S. cerevisiae*. Best performance is obtained by a weighted version of the TPR algorithm, followed by the TPR, a related hierarchical ensemble technique where information flows only from a node to its offspring, and a non-hierarchical ensemble technique that ignores the hierarchy in the ontology [157]. In another related work [188], a discrete approach is proposed that infers the most probable TPR-consistent assignments. The GO DAG is modeled as a Bayesian network that infers the most probable assignments via global optimization. The differential evolution algorithm is adapted for this purpose.

8.3 Kernel Integration Methods

Kernel-based methods [310] encode the pairwise similarity between proteins in a similarity matrix, also called a kernel matrix. This is a positive definite and symmetric matrix $K(x, y)$, where elements record the similarity between proteins x, y. Different kernel matrices can be defined to encode protein-pair similarities according to different data sources. For instance, when employing protein sequences, there are various options. BLAST E-values can be used to fill the entries of the kernel matrix. Alternatively, the spectrum/string kernel [203], motif kernel [26], and Pfam kernel can be used [115]. In the string kernel, a protein sequence x is represented by a vector $\phi(x)$ of frequencies of all k-mers, and then the inner product of two vectors $\phi(x), \phi(y)$ corresponding to two proteins x, y is taken to obtain $K(x, y)$.

For structured data sources, such as protein–protein interaction data, the random walk kernel [315] and diffusion kernels [183] are employed. Diffusion kernels encode similarities between the nodes of a network and are variants of $K = e^{-\beta L}$, where $\beta > 0$ is the parameter that quantifies the degree of diffusion, and L is the network Laplacian. A comprehensive evaluation in [227] shows that diffusion kernels give superior performance on function prediction, prioritizing genes related to a phenotype, and identifying false positives and false negatives from RNAi experiments. Work in [227] concludes that diffusion kernels should be the kernel of choice to measure network similarity over other similarity measures, such as direct neighbors and short path distance.

Kernel matrices corresponding to different data sources can be combined by carrying out basic algebraic operations such as addition, multiplication, or exponentiation. When addition is employed, the individual kernels can be weighted by fixed coefficients [196, 275], or by coefficients learned via semi-definite programming [195]. The latter can be computationally demanding, particularly on large and multiple data sets. In response, more efficient combination schemes have been proposed recently based on semi-infinite programming [326]. Whatever the strategy employed to combine individual kernels, the resulting kernel matrix can then be fed to popular classifiers, such as SVM or kNN.

8.4 Network Integration Methods

Instead of encoding protein similarities on different data sources via kernels, network methods encode similarities as edges connecting protein pairs in a network. For instance, sequence similarities between proteins can be estimated via BLAST E-value or other means (analogous to spectrum or string kernels) and encoded in a network connecting proteins of similar sequences. Co-expressed genes can also be connected by edges in a gene expression network. Similar ideas can be employed to encode similarities based on phylogenetic profile or phylogenetic tree in networks. These network-based representations of biological data can be as powerful as PPI or gene–gene interaction networks and can be simultaneously exploited to predict function.

Work in [152] focuses on a drawback of many network-based methods that ignore dependencies between interacting pairs and predict them independently of one another. In [152], relational Markov networks are employed to build a unified probabilistic model that allows predicting unobserved interactions concurrently. The model integrates various attributes and models measurement noise. In essence, PPI networks, interaction assay readouts, and other protein attributes are represented as random variables. Variable dependencies are modeled by joint distributions. Since a naive representation of the joint distribution requires a large number of parameters, relational Markov models are used instead. Improved performance is reported over related methods for predicting sub-cellular localization and interaction partners of the mediator complex.

The MAGIC method proposed in [348] integrates yeast PPI data from the General Repository of Interaction Datasets (GRID) [37], pairs of genes that have experimentally determined bindings sites for the same transcription factor, as extracted from The (SCPD) Promoter Database of *S. cerevisiae* [396], and gene expression data. Three separate gene–gene relationship matrices are constructed from each data source, with an entry encoding whether a particular gene pair has a functional relationship or not; 0 indicates lack of relationship, and a numeric value indicates confidence of putative relationship. Different algorithms are used on each data sources to obtain these relationships. For instance, gene expression data are subjected to different clustering algorithms, such as K-means, SOM, and hierarchical clustering, and each of these algorithms are nodes in a Bayesian network constructed for each gene–gene pair. SCPD data provide gene–gene pairs directly. These matrices are provided as input to Bayesian networks, one for each gene-gene pair. A network combine evidence from the different clusters to generate a posterior belief for whether its corresponding gene-gene pair has a functional relationship. MAGIC is reported to improve accuracy of the functional groupings compared with gene expression analysis alone [348].

Work in [255] integrates functional linkage graphs constructed from PPI and gene expression data. Functional linkage graphs are constructed to encode via edges evidence for functional similarity. These graphs are used in concert with categorical data, such as protein motif data, mutant phenotype data, and protein localization data, to make a final prediction. The categorical features of a query protein are used as random variables/nodes in a Bayesian network, together with annotated neighbors of a query protein in the functional linkage graphs. The posterior probability of the query annotated with a particular GO term is then calculated. This approach is employed to predict functions for yeast proteins. A cross validation setting shows that this integrated approach increases recall by 18 %, compared to using PPI data alone at the 50 % precision. The integrated predictor also outperforms each individual predictor. However, improvements in performance are not uniform and depend on the particular functional category predicted.

In [126], information present in metabolic networks and gene co-expression data is indirectly combined. A graph distance function is first defined on metabolic networks, and the function is combined with a correlation-based distance function for gene expression measurements. The resulting distance function is used to jointly cluster genes and network vertices via hierarchical clustering. The resulting clusters are shown to be interpretable in terms of biochemical network and gene expression data. A related, clustering-based method is proposed in [347]. Co-expression and PPI networks are separately evaluated by computing the probability of groups of genes to be correlated in the networks. The groups of correlated genes are found via super-paramagnetic hierarchical clustering.

Different data sources, such as PPI data, gene expression, phenotypic sensitivity, and transcription factor binding are integrated in [336] in a bipartite graph, with genes on one side of the graph, and their properties on the other. Biclustering algorithms based on combinatorial principles are then used to detect statistically significant subgraphs that correspond to functionally related genes. In [318], a clustering method based on learning a probabilistic model, referred to as a hidden modular random field. The relation between hidden variables represents a given gene network. The learning algorithm minimizes an energy function that considers network modularity. The method is shown to be highly sensitive for gene clustering and annotation of gene function.

Recent work in [231] proposes semi-supervised parametric neural models to combine different bio-molecular networks and predict protein functions. The models take into account the unbalance between annotated and unannotated proteins in the construction of the integrated network and in the final prediction of annotations for each functional class. Evaluation on full-genome and ontology-wide experiments on three eukaryotic organisms show that the UNIPred method proposed in [231] compares favorably with state-of-the-art methods, such as SW [248] and MS-kNN [194].

8.5 State of Data Integration Methods for Function Prediction

An increasing volume and diversity of biological data presents both opportunities and challenges for data integration methods. One such challenge that is relatively under-explored in data integration methods is how to explore topologies of multiple different networks. The majority of data integration methods for function prediction exploit PPI networks but largely ignore other important network data, such as gene-gene interaction networks and metabolomic interaction networks. Some work exists in this direction via methods that use random walk or diffusion processes to infer knowledge from all networks concurrently, though in the context of predicting disease interactions, disease gene associations, drug target interactions, and drug disease associations [61, 87, 120, 142, 354]. In this context, many issues regarding data integration that are currently under-pursued in function prediction methods are being addressed, such as noise, bias in data collection, concordant and discordant data sets, and scalability. A comprehensive review of data integration methods for disease- and drug-driven problems in molecular biology can be found in [110].

Data integration is identified as a key direction to improve the performance of function prediction methods. A detailed study in [357] pitches two different network-based function prediction approaches against each other, ensemble techniques that combine classifiers versus state-of-the-art classifiers that integrate various datasets. The study reports that a modest benefit of 17 % in the area under the ROC (AUROC) is obtained from ensemble techniques over the baseline classifiers. In contrast, data aggregation results in an 88 % improvement in mean AUROC. The study concludes that substantial evidence supports the view that additional algorithm development has little to offer for gene function prediction as opposed to data aggregation. While a saturation point may have been reached for off-the-shelf machine learning methods, there may be further ground to explore for novel methods capable of efficiently and effectively integrating noisy and non-uniformly dense data.

9 Text Mining-Based Methods for Function Prediction

Text mining is a promising machine learning technology for the analysis of biomedical literature in the problem of protein function classification for the fundamental reason of the abundance of literature that links proteins with each other. This offers the hope of increasing the size of labeled data available for training and evaluation.

Following the idea of using query proteins to find their homologous proteins, one of the first applications of text mining was to utilize this notion. In Renner and Aszodi [294] the authors describe a procedure for the prediction of functions of novel products. The last steps of the procedure are based on text mining. First, a protein whose function remains unknown is used as input across multiple databases (e.g., SWISS-PROT, PIR, PROSITE). The result of these searches are annotation documents that can be subjected to text mining procedures. In particular, the documents are compared by checking the terms that occur in them and using those terms to produce clusters. The principle is that if two documents contain terms that belong to the same cluster, then the documents probably describe the same phenomenon. In order to cluster the terms, the authors analyze their co-occurrence in documents, proceeding to build clusters starting at a term and adding terms that often co-occur with it, recursively. The probability of a term belonging to a cluster is then computed as the ratio of the sum of the number of times the term co-occurred with all the other terms in the same cluster, divided by the total number of occurrences of the term. For a given document, the "match" score of a cluster is defined as the maximum probability of belonging to that cluster for all the terms found in the document. Comparing documents is now a matter of computing the normalized sum of differences of their match scores across all the clusters. With such a distance measure, the documents can be clustered. (While this is a valid way to compute distances, it is intriguing to think what the results would have been if a more modern way of clustering documents, such as Topic Modeling [31] had been utilized.) It was observed that for most proteins, all the documents clustered into a single clustering, indicating coherence.

Simple text classification approaches have been used for the prediction of functionality, using the documents with which the proteins are associated. Raychaudhari [291] uses the maximum entropy, naive Bayes, and nearest neighbor classifiers using training abstracts from PubMed. The features for the classifiers are bigrams of two co-occurring words subjected to a $\tilde{\chi}^2$ test of correlation with the class. Results show that the most accurate of classifiers, namely maximum entropy, is capable of finding the proper class 72.8 % of the time.

A kNN classifier is applied by Keck and Wetter [166] using BLAST searches and a variety of databases (GenProtEC, MIPS). Results show a very low recall of 0.4, which can be attributed to a very weak distance metric.

A more elaborate approach of integrating text mining into this problem can be found in the work of Eskin and Agichtein [97]. In that paper, the authors use the SVM classifier, combining a variety of text and sequence kernels. First, a seed set set is created. The set consists of labeled proteins as positive examples and other proteins with different labels as negative. This set is small due to the availability of known cases. A text classifier is then trained over the annotations of the sequences in the set (found in a variety of databases). The feature space employed is the bag-of-words representation of the annotations. Each word in the annotation receives a 1 for the word's dimension; words not present receive a 0. This results in a feature space that is high-dimensional but very sparse. The text kernel is the dot product of the representation vectors. This classifier, after being trained with the original set, is used to predict the function of unknown proteins, using textual information available in databases such as SWISS-PROT. The result of applying the trained classifier is an enriched, larger set of labeled proteins.

Next, a joint classifier that uses sequence information and text annotations is trained with the new set. To this end, a kernel for both sequences and text is defined. The text part of the kernel is, as described above, based on the bag-of-words representation. The sequence part of the kernel represents sequences as substrings of length k, or k-mers, obtained by segmenting the sequence with a moving window. The feature space contains a dimension for each possible k-mer, with a 1 for a k-mer that appears in the sequence and 0s for those that do not. The kernel is then the inner product of two such representations. Since matching k-mers in practice are very rare, the authors utilize the sparse kernel representation that allows for approximate matching. In it, the kernel is defined as a parameter α raised to the Hamming distance between the sequences being compared. The combination of the text and sequence kernels is achieved by kernel composition. The result is a kernel that adds the two components and a degree two polynomial kernel over the sum of the two original kernels. The rationale for the polynomial term is to include features for all pairs of sequences and words. With such combined kernel, the classifier effectively learns from both sequence and text annotations and the interactions between them.

An additional benefit of this approach is obtained by projecting the classifier onto the original sequences to learn which regions of the protein have a high positive weight with respect to the class and as such, are likely candidates for relevant functional regions. Experiments conducted by leaving 20 % of the set as a test set and using the remaining as the training set indicate that the results obtained by the joint classifier are superior across a variety of functional classes to those obtained by applying any of the two (text or sequence) classifiers independently. The task of identifying relevant regions is shown to perform well when comparing the results to a searchable database, such as NLS [253].

10 Discussion and Prospects for Future Research

A pervasive theme of this survey of function prediction methods has been that while significant advances are occurring in each of the five categories of methods for function prediction, significant performance gains are obtained by methods that ingrate data from diverse sources. Two recent studies point to the fact that data integration is expected indeed to be the most promising avenue for improved function prediction performance [231, 357]. Readers with interests beyond computational protein function prediction may find useful information in this survey on how data are integrated in the machine learning methods summarized here. Interesting trends can be observed regarding how different types of features are combined, and how such trends have evolved over time as driven by the need to balance between accuracy and computational efficiency.

Considering all the rapid advancements in novel methodologies for function prediction, it is not easy to keep track of the current state of automated function prediction. Nor is it easy to objectively conclude whether certain methods are better than others from summaries of published works, where performance is evaluated in a controlled setting and on some specific dataset of interest. The CAFA experiment provides just the avenue for objective comparisons. A large-scale evaluation of 54 automated function prediction methods in CAFA is reported in [287]. Two main findings are reported: first, that current methods significantly outperform first-generation ones on all types of query proteins; second, that, although current methods perform well enough to guide experiments, there is significant room for improvement.

Specifically, the top five labs/methods in 2013 CAFA on all targets, one- and multiple-domain proteins are Jones-UCL [70], GOstruct [324], Argot2 [98], ConFunc [368], and PANNZER [186]. Their comparative performance is summarized in Fig. 5. The Jones-UCL team consistently outperformed other methods due to a massive integration of evolutionary analyses and multiple data sources, combining in a probabilistic manner GO term predictions from PSI-BLAST, SWISS-PROT text mining, amino-acid trigram mining, FFPred sequence features [225], orthologous groups, PSSM profile–profile comparisons, and FunctionSpace [224]. A network propagation algorithm based on the GO graph structure combines the various predictions. ConFunc [368], we recall, is another data integration method. GOstruct models the structure of the GO hierarchy in the framework of kernel methods for

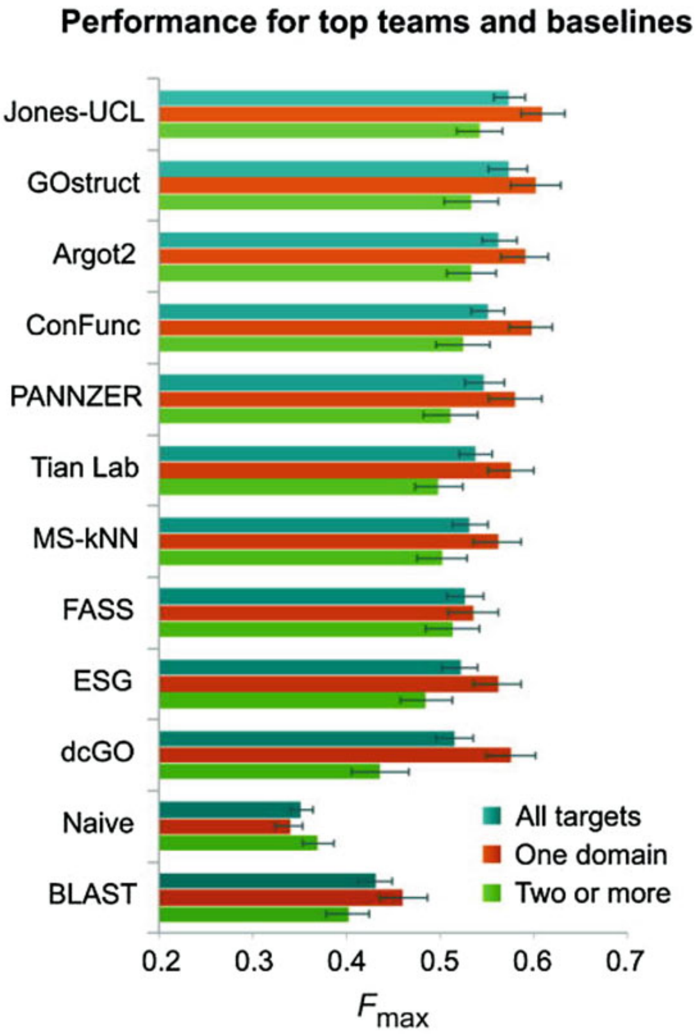


Fig. 5 The performance of the top ten methods is shown here. The methods achieved higher accuracy on single-domain proteins. Confidence intervals (95 %) were determined using bootstrapping with 10,000 iterations on the target sequences

structured-output spaces. The structured output SVM in [324] does not only well in CAFA 2013, but it also confers high performance to approaches that investigate text-mined features for automated function prediction with GOstruct [105].

In contrast, The Argot2 web server debuted in [98] is mainly a sequence-based method, that employs BLAST and HMMER searches of a query sequence against UnitProtKB and PFAM databases. GO terms are weighted based on E-values returned by the searches, and the weights are processed according to semantic similarity relations between the terms. A recent version of the web server, Argot2.5, is enriched with more features, including a new semantic similarity measure, and shown to improve performance even further over Argot2 [197]. Argot2.5 is also shown to outperform PANNZER [186], which is another sequence-based method. PANNZER relies on a weighted k-nearest neighbor approach with statistical testing, after partitioning sequence-similarity results into clusters according to description similarity. A sophisticated regression model evaluates the support for the candidate cluster.

The list of the top performers is diverse in terms of methodologies. Perhaps not surprisingly, and in line with other studies drawn similar observations [124], sophisticated sequence-based methods can perform comparably to state-of-the-art data integration methods. A pre-print of a recent, 2016 report can also be found [158]. In the report, 126 methods from 56 research groups are compared against one another in a set of 3681 proteins from 18 species. One of the findings in the report is that top-performing methods in CAFA2 outperform top-performing ones in CAFA1. This finding suggests that indeed computational function prediction is improving, possibly due to both an increase in experimental annotations (via high-throughout wet-laboratory techniques) and improvements in methodology.

Carefully drawn case studies in [287] show that there is room for improvement. One challenge in function prediction that is not often mentioned is related to the existence of promiscuous proteins that are multi-functional; indeed, more than 30 % of the proteins in SWISS-PROT have more than one leaf in the Molecular Function ontology; more than 60 % have more than one leaf in the Biological Process ontology. In addition, while data integration is often touted as the most promising direction, the presence of noisy and erroneous experimental data may be an additional source of error that needs to be addressed for robust performance. Finally, machine learning is shown to generally improve performance, and it is expected that there is more performance to be gained by approaches based on principles of statistical learning and inference.

Acknowledgements Funding for this work is provided in part by NSF-IIS1144106.

References

1. Abascal, F., Valencia, A.: Automatic annotation of protein function based on family identification. Proteins **53**(3), 683–692 (2003)
2. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD Intl Conf on Management of Data, pp. 207–216. ACM (1993)
3. Albert, R.: Network inference, analysis, and modeling in systems biology. Plant Cell **19**(11), 3327–3338 (2007)
4. Alberts, B., Johnson, A., Lewis, J., et al.: From RNA to protein. In: Molecular Biology of the Cell, 4 edn. New York: Garland Science (2002)
5. Alberts, B., Johnson, A., Lewis, J., et al.: Studying gene expression and function. In: Molecular Biology of the Cell, 4 edn. New York: Garland Science (2002)
6. Alexandrov, N.N.: SARFing the PDB. Protein Eng **9**(9), 727–732 (1996)
7. Altman, D.G.: Practical Statistics for Medical Research. Chapman and Hall (1997)

8. Altschul, S.F., Madden, T.L., Schaeffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. **25**, 3389–3402 (1997)
9. Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., Murzin, A.G.: Scop database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res **32**(Database issue), D226–D229 (2004)
10. Apeltsin, L., Morris, J.H., Babbitt, P.C., Ferrin, T.E.: Improving the quality of protein similarity network clustering algorithms using the network edge weight distribution. Bioinformatics **27**(3), 326–333 (2011)
11. Arnau, V., Mars, S., Marin, I.: Iterative cluster analysis of protein interaction data. Bioinformatics **21**(3), 364–378 (2005)
12. Ashburner, M., Ball, C., Blake, K., et al.: The gene ontology consortium. Nature Genetics **25**(1), 25–29 (2000)
13. Aung, Z., Tan, K.L.: Rapid 3D protein structure database searching using information retrieval techniques. Bioinformatics **20**(7), 1045–1052 (2004)
14. Badea, L.: Functional discrimination of gene expression patterns in terms of the gene ontology. In: Pacific Symp Biocomput (PSB), pp. 565–576 (2003)
15. Bader, G.D., Betel, D., Hogue, W.V.: BIND: the biomolecular interaction network database. Nucleic Acids Res **31**(1), 248–250 (2003)
16. Bailey, T.L., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Intl Conf Intell Sys Mol Biol (RECOMB), pp. 28–36 (1998)
17. Bailey, T.L., Gribskov, M.: Combining evidence using p-values: application to sequence homology searches. Bioinformatics **14**(1), 48–54 (1998)
18. Bairoch, A., BUcher, P., Hoffmann, K.: The PROSITE database, its status in 1997. Nucl. Acids Res. **25**(1), 217–221 (1997)
19. Bar-Joseph, Z.: Analyzing time series gene expression data. Bioinformatics **20**(16), 2493–2503 (2004)
20. Bar-Joseph, Z., Gitter, A., Simon, I.: Studying and modelling dynamic biological processes using time-series gene expression data. Nat Rev Genet **13**(8), 552–564 (2012)
21. Barabasi, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nature Rev Genet **5**(2), 101–113 (2004)
22. Barrett, et al.: NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res **41**(Database issue), D991–D995 (2013)
23. Bder, G., Hogue, C.: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinf **4**(1), 2 (2003)
24. Bellaachia, A., Portnov, D., Chen, Y., Elkahloun, A.G.: E-CAST: a data mining algorithm for gene expression data. In: Workshop on Data Mining in Bioinformatics (BIOKDD), pp. 49–54 (2002)
25. Ben-Dor, A., Shamir, R., Yakhini, Z.: Clustering gene expression patterns. J Comput Biol **6**(3–4), 281–297 (1999)
26. Ben-Hur, A., Brutlag, D.: Remote homology detection: a motif based approach. Bioinformatics **19**(Suppl 1), i26–i33 (2003)
27. Ben-Hur, A., Brutlag, D.: Sequence motifs: Highly predictive features of protein function. In: I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh (eds.) Feature extraction and foundations and applications. Springer Verlag (2005)
28. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., , Bourne, P.E.: The protein data bank. Nucl. Acids Res. **28**(1), 235–242 (2000)
29. Bilu, Y., Linial, M.P.: Functional consequences in metabolic pathways from phylogenetic profiles. In: Intl Workshop on Algorithms in Bioinformatics (WABI), pp. 263–276 (2002)
30. Blatt, M., Wiseman, S., Domany, E.: Superparamagnetic clustering of data. FEBS Lett **76**, 3251–3254 (1996)
31. Blei, D.: Probabilistic topic models. Communications of the ACM **55**(4), 77–84 (2012)
32. Blei, D.M.: Latent Dirichlet Allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

33. Blekas, K., Fotiadis, D.I., Likas, A.: Motif-based protein sequence classification using neural networks. J Comput Biol **12**(1), 64–82 (2005)
34. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res **31**(1), 365–370 (2003)
35. Bork, P., Koonin, E.V.: Protein sequence motifs. Curr Opin Struct Biol **6**(3), 366–376 (1996)
36. Braberg, H., Webb, B.M., Tjioe, E., Pieper, U., Sali, A., Madhusudhan, M.S.: SALIGN: a web server for alignment of multiple protein sequences and structures. Bioinformatics **15**(28), 2071–2073 (2012)
37. Breitkreutz, B., Stark, C., Tyers, M.: The GRID: The general repository for interaction datasets. Genome Biol **4**(3), R3 (2003)
38. Brenner, S.E.: Errors in genome annotation. Trends Genet **15**(4), 132–133 (1999)
39. Brenner, S.E., Levitt, M.: Expectations from structural genomics. Protein Sci. **9**(1), 197–200 (2000)
40. Brown, K.R., Jurisica, I.: Online predicted human interaction database. Bioinformatics **21**(9), 2076–2082 (2005)
41. Brown, M.P., et al.: Knowledge based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA **97**(1), 262–267 (2000)
42. Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guénoche, A., Jacq, B.: Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol **5**(1), R6 (2003)
43. Bryan, K., Cunningham, P., Bolshakova, N.: Biclustering of expression data using simulated annealing. In: IEEE Symp Computer-based Medical Systems (CBMS), pp. 383–388 (2005)
44. Bucak, S., Jin, R., Jain, A.: Multi-label multiple kernel learning by stochastic approximation: Application to visual object recognition. In: Advances Neural Inform Processing Systems (NIPS), pp. 1145–1154 (2010)
45. Budowski-Tal, I., , Nov, Y., Kolodny, R.: Fragbag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. Proc. Natl. Acad. Sci. USA **107**, 3481–3486 (2010)
46. Butte, A.J., Bao, L., Reis, B.Y., Watkins, T.W., Kohane, I.S.: Comparing the similarity of time-series gene expression using signal processing metrics. J Biomed Bioinf **34**(6), 396–405 (2001)
47. Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., Chen, Y.Z.: SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res **31**(13) (2003)
48. Cai, Y.D., Doig, A.J.: Prediction of saccharomyces cerevisiae protein functional class from functional domain composition. Bioinformatics **20**(8), 1292–1300 (2004)
49. Califano, A.: SPLASH: structural pattern localization analysis by sequential histograms. Bioinformatics **16**(4), 341–357 (2000)
50. Cao, R., Cheng, J.: Integrated protein function prediction by mining function associations, sequences, and protein-protein and gene-gene interaction networks. Methods **93**, 84–99 (2016)
51. Carpentier, M., Brouillet, S., Pothier, J.: YAKUSA: a fast structural database scanning method. Proteins: Struct. Funct. Bioinf. **61**(1), 137–151 (2005)
52. Carugo, O.: Rapid methds for comparing protein structures and scanning structure databases. Current Bioinformatics **1**, 75–83 (2006)
53. Carugo, O., Pongor, S.: Protein fold similarity estimated by a probabilistic approach based on c(alpha)-c(alpha) distance comparison. J Mol Biol **315**(4), 887–898 (2002)
54. Chakrabarti, S., Venkatramanan, K., Sowdhamini, R.: SMoS: a database of structural motifs of protein superfamilies. Protein Eng **16**(11), 791–793 (2003)
55. Chatr-Aryamontri, A., et al.: The BioGRID interaction database: 2015 update. Nucleic Acids Res **43**(Database Issue), D470–D478 (2015)
56. Chen, C., Chung, W., Su, C.: Exploiting homogeneity in protein sequence clusters for construction of protein family hierarchies. Pattern Recognition **39**(12), 2356–2369 (2006)

57. Chen, L., Xuan, J., Riggins, R.B., Wang, Y., Clarke, R.: Identifying protein interaction subnetworks by a bagging markov random field-based method. Nucleic Acd Res **41**(2), e42 (2013)
58. Chen, Y.J., Kodell, R., Sistare, F., Thompson, K.L., Moris, S., Chen, J.J.: Studying and modelling dynamic biological processes using time-series gene expression data. J Biopharm Stat **13**(1), 57–74 (2003)
59. Chen, Y.J., Mamidipalli, S., Huan, T.: HAPPI: an online database of comprehensive human annotated and predicted protein interactions. BMC Genomics **10**(Suppl 1), S16 (2009)
60. Cheng, B.Y., Carbonell, J.G., Klein-Seetharaman, J.: Protein classification based on text document classification techniques. Proteins **58**(4), 955–970 (2005)
61. Cheng, F., et al.: Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol **8**(5), e1002,503 (2012)
62. Cheng, Y., Church, G.M.: Biclustering of expression data. In: Intl Conf Intell Sys Mol Biol (RECOMB), pp. 93–103 (2000)
63. Chitale, M., Hawkins, T., Park, C., Kihara, D.: ESG: extended similarity group method for automated protein function prediction. Bioinformatics **25**(14), 1739–1745 (2009)
64. Cho, Y., Zhang, A.: Predicting protein function by frequent functional association pattern mining in protein interaction networks. IEEE Trans Info Technol Biomed **14**(1), 30–36 (2009)
65. Chua, H.N., Sung, W.K., Wong, L.: Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics **22**(13), 1623–1630 (2006)
66. Clark, W.T., Radivojac, P.: Analysis of protein function and its prediction from amino acid sequence. Proteins: Struct Funct Bioinf **79**(7), 2086–2096 (2011)
67. Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marrafini, L.A., Zhang, F.: Multiplex genome engineering using CRISPR/Cas systems. Science **339**(6121), 819–823 (2013)
68. Consortium, T.U.: Ongoing and future developments at the universal protein resource. Nucleic Acids Res **39**(Database issue), D214–D219 (2011)
69. Cowley, M.J., Pinese, M., Kassahn, K.S., Waddell, N., Pearson, J.V., Grimmond, S.M., Biankin, A.V., Hautaniemi, S., Wu, J.: PINA v2.0: mining interactome modules. Nucleic Acids Res **40**(Database issue), D862–D865 (2012)
70. Cozzetto, D., Buchan, D.W.A., Jones, D.T.: Protein function prediction by massive integration of evolutionary analyses and multiple data sources. BMC Bioinf **14**(Suppl 1), S1 (2013)
71. Dandekar, T., Snel, B., Huynen, M., Bork, P.: Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci **23**(9), 324–328 (1998)
72. Das, R., Kalita, J., Bhattacharyya, D.K.: A new approach for clustering gene expression time series data. Intl J Bioinform Res Appl **5**(3), 310–328 (2009)
73. Date, S.V., Marcotte, E.M.: Protein function prediction using the Protein Link EXplorer (PLEX). Bioinformatics **21**(10), 2558–2559 (2005)
74. Déjean, S., Martin, P.G.P., Besse, P.: Clustering time-series gene expression data using smoothing spline derivatives. EURASIP J Bioinf Sys Biol **2007**(1), 70,561 (2007)
75. Deng, M., Sun, T., Chen, T.: Assessment of the reliability of protein-protein interactions and protein function prediction. In: Pacific Symp Biocomput (PSB), vol. 8, pp. 140–151 (2003)
76. Deng, M., Tu, Z., Sun, F., Chen, T.: Mapping gene ontology to proteins based on protein-protein interaction data. Bioinformatics **20**(6), 895–902 (2004)
77. Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F.: Prediction of protein function using protein-protein interaction data. J Comput Biol **10**(6), 947–960 (2003)
78. Deng, X., Ali, H.H.: A hidden markov model for gene function prediction from sequential expression data. In: IEEE Comput Sys Bioinf Conf (CSB), pp. 670–671 (2004)
79. Devos, D., Valencia, A.: Practical limits of function prediction. Proteins: Struct Funct Bioinf **41**(1), 98–107 (2000)
80. Doerks, T., Bairoch, A., Bork, P.: Protein annotation: detective work for function prediction. Trends Genet **14**(6), 248–250 (1998)
81. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2 edn. Wiley-Interscience (2000)

82. Dwight, S.S., et al.: Saccharomyces genome database (SGD) provides secondary gene annotation using the gene ontology (GO). Nucleic Acids Res **30**(1), 69–72 (2002)
83. Eddy, S.R.: Profile hidden Markov models. Bioinformatics **14**(9), 755–763 (1998)
84. Edgar, R., Domrachev, M., Lash, A.E.: Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res **30**(1), 207–210 (2003)
85. Eisen, J.A.: Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res **8**(3), 163–167 (1998)
86. Eisner, R., , Poulin, B., Szafron, D., Lu, P., Greiner, R.: Improving protein function prediction using the hierarchical structure of the gene ontology. In: IEEE Comput Intell Bioinf Comput Biol (CIBCB), pp. 1–8 (2005)
87. Emig, D., Ivliev, A., Pustovalova, O., Lancashire, L., Bureeva, S., Nikolsky, Y., Bessarabova, M.: Drug target prediction and repositioning using an integrated network-based approach. PLoS One **8**(4), e60,618 (2013)
88. Enault, F., Suhre, K., Abergel, C., Poirot, O., Claverie, J.: Annotation of bacterial genomes using improved phylogenomic profiles. Bioinformatics **19**(Suppl 1), i105–i107 (2003)
89. Enault, F., Suhre, K., Abergel, C., Poirot, O., Claverie, J.: Phydbac (phylogenomic display of bacterial genes): An interactive resource for the annotation of bacterial genomes. Nucleic Acids Res **31**(13), 3720–3722 (2003)
90. Enault, F., Suhre, K., Abergel, C., Poirot, O., Claverie, J.: Phydbac2: improved inference of gene function using interactive phylogenomic profile and chromosomal location analysis. Nucleic Acids Res **32**(Web Server Issue), W336–W339 (2004)
91. Enault, F., Suhre, K., Claverie, J.: Phydbac "gene function predictor": a gene annotation tool based on genomic context analysis. BMC Bioinf **6**(247) (2005)
92. Engelhardt, B.E., Jordan, M.I., Muratore, K.E., Brenner, S.E.: Protein molecular function prediction by bayesian phylogenomics. PLoS Comput Biol **1**(5), e45 (2005)
93. Enright, A.J., Ouzounis, C.A.: Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol **2**(9), RESEARCH0034 (2001)
94. Enright, A.J., Van Dongen, S., Ouzounis, C.A.: An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res **30**(7), 1575–1584 (2002)
95. Erickson, H.P.: Cooperativity in protein-protein association: the structure and stability of the actin filament. J Mol Biol **206**(3), 465–474 (1989)
96. Ernst, J., Nau, G.J., Bar-Joseph, Z.: Clustering short time series gene expression data. Bioinformatics **21**(Suppl 1), i159–i168 (2005)
97. Eskin, E., Agichtein, E.: Combining text mining and sequence analysis to discover protein functional regions. In: Pac. Symp. Biocomputing, pp. 288–299 (2004)
98. Falda, M., et al.: Argot2: a large scale function prediction tool relying on semantic similarity of weighted gene ontology terms. BMC Bioinf **28**(Suppl 4), S14 (2012)
99. Fayech, S., Essoussi, N., Limam, M.: Partitioning clustering algorithms for protein sequence data sets. BioData Mining **2**(1), 3 (2009)
100. Felsenstein, J.: PHYLIP - phylogeny inference package (version 3.2). Cladistics **5**, 164–166 (1989)
101. Ferrer, L., Dale, J.M., Karp, P.D.: A systematic study of genome context methods: calibration, normalization and combination. BMC Genomics **11**(1), 1–24 (2010)
102. Fetrow, J.S., Siew, N., Di Gennaro, J.A., Martinez-Yamout, M., Dyson, H.J., Skolnick, J.: Genomic-scale comparison of sequence- and structure-based methods of function prediction: Does structure provide additional insight? Protein Science : A Publication of the Protein Society **10**(5), 1005–1014 (2001)
103. Forslund, K., Sonnhammer, E.L.: Predicting protein function from doma in content. Bioinformatics **24**(15), 1681–1687 (2008)
104. French, L.: Fast protein superfamily classification using principal component null space analysis. appendix a: A survey on remote homology detection and protein superfamily classification. Master's thesis, University of Windsor, Ontario, Canada (2005)
105. Funk, C.S., Kahanda, I., Ben-Hur, A., Verspoor, K.M.: Evaluating a variety of text-mined features for automatic protein function prediction with GOstruct. J Biomed Semantics **18**(6), 9 (2015)

106. Gascuel, O.: BIONJ: an improved version of the nj algorithm based on a simple model of sequence data. Mol Biol Evol **14**(7), 685–695 (1997)
107. Gether, U.: Uncovering molecular mechanisms involved in activation of g protein-coupled receptors. Endocr Rev **21**(1), 90–113 (2000)
108. Gibrat, J.F., Madej, T., Bryant, S.H.: Surprising similarities in structure comparison. Curr. Opinion Struct. Biol. **6**(3), 377–385 (1996)
109. Gillis, J., Pavlidis, P.: The role of indirect connections in gene networks in predicting function. Bioinformatics **27**(13), 1860–1866 (2011)
110. Gligorijevic, V., Przulj, N.: Methods for biological data integration: perspectives and challenges. Roy Soc Interface **12**(112), 20150,571 (2015)
111. Godzik, A., Skolnick, J.: Flexible algorithm for direct multiple alignment of protein structures and sequences. Comput Appl Biosci **10**(6), 587–596 (1994)
112. Goh, C., Bogan, A.A., Joachimiak, M., Walther, D., Cohen, F.E.: Co-evolution of proteins with their interaction partners. J Mol Biol **299**(2), 283–293 (2000)
113. Goldberg, D.S., Roth, F.P.: Assessing experimentally derived interactions in a small world. Proc Natl Acad Sci USA **100**(8), 4372–4376 (2003)
114. Goll, J., Rajagopala, S.V., Shiau, S.C., Wu, H., Lamb, B.T., Uetz, P.: MPIDB: the microbial protein interaction database. Bioinformatics **24**(15), 1743–1744 (2008)
115. Gomez, S.M., Noble, W.S., Rzhetsky, A.: Learning to predict protein-protein interactions from protein sequences. Bioinformatics **19**(15), 1875–1881 (2003)
116. Gong, Q., Ning, W., Tian, W.: GoFDR: A sequence alignment based method for predicting protein functions. Methods **S1046–2023**(15), 30,048–7 (2015)
117. Guan, Y., Myers, C.L., Hess, D.C., Barutcuoglu, Z., Caudy, A.A., Troyanskaya, O.G.: Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biol **9**(Suppl 1), S3 (2008)
118. Gui, J., Li, H.: Mixture functional discriminant analysis for gene function classification based on time course gene expression data. In: Joint Statistical Meeting: Biometrics Section (2003)
119. Gúldener, U., Muensterkoetter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H.W., Stúmpflen, V.: MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res **34**(Database issue), D436–D441 (2006)
120. Guo, X., Gao, L., Wei, C., Yang, X., Zhao, Y., Dong, A.: A computational method based on the integration of heterogeneous networks for predicting disease-gene associations. PLoS One **6**(e24171) (2011)
121. Guruprasad, K., Prasad, M.S., Kumar, G.R.: Database of structural motifs in proteins. Bioinformatics **16**(4), 372–375 (2000)
122. Guthke, R., Schmidt-Heck, W., Hahn, D., Pfaff, M.: Gene expression data mining for functional genomics. In: European Symp Intelligent Techniques, pp. 170–1777 (2000)
123. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach Learn **46**(1–3), 389–422 (2002)
124. Hamp, T., et al.: Homology-based inference sets the bar high for protein function prediction. BMC Bioinf **14**(Suppl 1), S7 (2013)
125. Han, L.Y., Zheng, C.J., Lin, H.H., Cui, J., Li, H., Zhang, H.L., Tang, Z.Q., Chen, Y.Z.: Prediction of functional class of novel plant proteins by a statistical learning method. New Phytol **168**(1), 109–121 (2005)
126. Hanisch, D., Zien, A., Zimmer, R., Lengauer, T.: Co-clustering of biological networks and gene expression data. Bioinformatics **18**(Suppl 1), S145–S154 (2002)
127. Hartigan, J.A.: Direct clustering of a data matrix. J Amer Stat Assoc **67**(337), 123–129 (1972)
128. Hartuv, E., Shamir, R.: A clustering algorithm based on graph connectivity. Information Processing Letters **76**(4–6), 175–181 (2000)
129. Hawkins, T., Chitale, M., Luban, S., Kihara, D.: PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins: Struct Funct Bioinf **74**(3), 566–582 (2009)
130. Hawkins, T., Luban, S., Kihara, D.: Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci **15**(6), 1550–1556 (2006)

131. Hayete, B., Bienkowska, J.R.: GOTrees: Predicting go associations from protein domain composition using decision trees. In: Pacific Symp Biocomput (PSB), pp. 140–151 (2005)
132. Heard, N., Holmes, C.C., Stephens, D.A., Hand, D.J., Dimopoulos, G.: Bayesian coclustering of anopheles gene expression time series: Study of immune defense response to multiple experimental challenges. Proc Natl Acad Sci USA **102**(47), 16,939–16,944 (2005)
133. Hegyi, H., Gerstein, M.: The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol **288**(1), 147–164 (1999)
134. Hinson, J.T., Chopra, A., Nafissi, N., Polacheck, W.J., Benson, C.C., Swist, S., Gorham, J., Yang, L., Schafer, S., Sheng, C.C., Haghighi, A., Homsy, J., Hubner, N., Church, G., Cook, S.A., Linke, W.A., Chen, C.S., Seidman, J.G., Seidman, C.E.: Heart disease. titin mutations in iPS cells define sarcomere insufficiency as a cause of dilated cardiomyopathy. Science **349**(6251), 892–986 (2015)
135. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T.: Assessment of prediction accuracy of protein function from protein-protein interaction data. Yeast **18**(6), 523–531 (2001)
136. Holm, L., Sander, C.: Protein structure comparison by alignment of distance matrices. jmb **233**(1), 123–138 (1993)
137. Hou, J., Chi, X.: Predicting protein functions from PPI networks using functional aggregation. Mathematical Biosciences **240**(1), 63–69 (2012)
138. Hou, J., S.-R., J., Zhang, C., Kim, S.: Global mapping of the protein structure space and application in structure-based inference of protein function. Proc. Natl. Acad. Sci. USA **102**, 3651–3656 (2005)
139. Hou, Y., Hsu, W., Lee, M.L., Bystroff, C.: Efficient remote homology detection using local structure. Bioinformatics **19**(17), 2294–2301 (2003)
140. Hsu, P.D., Lander, E.S., Zhang, F.: Development and applications of CRISPR-Cas9 for genome engineering. Cell **157**(6), 1262–1278 (2014)
141. Huang, J.Y., Brutlag, D.L.: The EMOTIF database. Nucleic Acids Res **29**(1), 202–204 (2001)
142. Huang, Y., Yeh, H., Soo, V.: Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. BMC Med Genomics **6**(3), S4 (2013)
143. Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P., Bairoch, A.: Recent improvements to the PROSITE database. Nucl. Acids Res. **32**(1), D134–D137 (2003)
144. Hulo, N., et al.: The PROSITE database. Nucleic Acids Res **34**(Database issue), D227–D230 (2006)
145. Humphrey, W., Dalke, A., Schulten, K.: VMD - Visual Molecular Dynamics. J. Mol. Graph. Model. **14**(1), 33–38 (1996). http://www.ks.uiuc.edu/Research/vmd/
146. Hunter, S., et al.: InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res **40**(Database issue), 306–312 (2012)
147. Huynen, M., Snel, B., Lathe, W., Bork, P.: Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res **10**(8), 1204–1210 (2000)
148. Hvidsten, T., Komorowski, J., Sandvik, A., Laegreid, A.: Predicting gene function from gene expressions and ontologies. In: Pacific Symp Biocomput (PSB), pp. 299–310 (2001)
149. Iakoucheva, L.M., Dunker, A.K.: Order, disorder, and flexibility: Prediction from protein sequence. Structure **11**(11), 1316–1317 (2003)
150. Jaakkola, T., Diekhans, M., Haussler, D.: Using the fisher kernel method to detect remote protein homologies. In: T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.W. Mewes, R. Zimmer (eds.) Int Conf Intell Sys Mol Biol (ISMB), pp. 149–159. AAAI Press, Menlo Park, CA (1999)
151. Jaakkola, T., Diekhans, M., Haussler, D.: A discriminative framework for detecting remote protein homologies. J Comput Biol **7**(1–2), 95–114 (2000)
152. Jaimovich, A., Elidan, G., Margalit, H., Friedman, N.: Towards an integrated protein-protein interaction network: A relational markov network approach. J Comput Biol **13**(2), 145–164 (2006)
153. Jensen, L., et al.: Prediction of human protein function from post-translational modifications and localization features. J Mol Biol **319**(5), 1257–1265 (2002)

154. Jensen, L.J., Gupta, R., Staerfeldt, H., Brunak, S.: Prediction of human protein function according to gene ontology categories. Bioinformatics **19**(5), 635–642 (2003)
155. Jiang, D., Pei, J., Ramanathan, M., Tang, C., Zhang, A.: Mining coherent gene clusters from gene-sample-time microarray data. In: ACM Intl Conf Knowledge Discovery Data Mining (SIGKDD), pp. 430–439 (2004)
156. Jiang, J.Q.: Learning protein functions from bi-relational graph of proteins and function annotations. In: Algorithms in Bioinformatics, *Lecture Notes in Computer Science*, vol. 6833, pp. 128–138. Springer Verlag (2011)
157. Jiang, X., Nariai, N., Steffen, M., Kasif, S., Kolaczyk, E.: Integration of relational and hierarchical network information for protein function prediction. BMC Bioinf **9**, 350 (2008)
158. Jiang, X., et al.: An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Quantitative Methods arXiv pp. 1–70 (2016)
159. Joshi, T., Xu, D.: Quantitative assessment of relationship between sequence similarity and function similarity. BMC Genomics **8**(1), 1–10 (2007)
160. Kabsch, W.: Efficient remote homology detection using local structure. Acta. Crystallog. sect. A **34**, 827–828 (1978)
161. Kalathur, R.K., Pinto, J.P., Hernández-Prieto, M.A., Machado, R.S., Almeida, D., Chaurasia, G., Futschik, M.E.: UniHI 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks. Nucleic Acids Res **42**(Database issue), D408–D414 (2014)
162. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M.: The KEGG resource for deciphering the genome. Nucleic Acids Res **32**(Database Issue), D277–D280 (2004)
163. Karaoz, U., Murali, T.M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C.R., Kasif, S.: Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci USA **101**(9), 2888–2893 (2004)
164. Karplus, K., Barret, C., Hughey, R.: Hidden markov models for detecting remote protein homologies. Bionformatics **14**(10), 846–856 (1998)
165. Keasar, C., Kolodny, R.: Using protein fragments for searching and data-mining protein databases. In: AAAI Workshop, pp. 1–6 (2013)
166. Keck, H., Wetter, T.: Functional classification of proteins using a nearest neighbor algorithm. In Silico Biology **3**(3), 265–275 (2003)
167. Kelley, L.A., Sternberg, M.J.: rotein structure prediction on the web: a case study using the phyre server. Nat Protocols **4**(3), 363–371 (2009)
168. Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M., D., K.P.: EcoCyc: a comprehensive database resource for escherichia coli. Nucleic Acids Res **33**(Database Issue), D334–D337 (2005)
169. Keshava, P., et al.: Human protein reference database–2009 update. Nucleic Acids Res **37**(Database issue), D767–D772 (2009)
170. Khan, I., Wei, Q., Chapman, S., Dukka, B.K., Kihara, D.: The PFP and ESG protein function prediction methods in 2014: effect of database updates and ensemble approaches. GigaScience **4**, 43 (2015)
171. King, A., Przulj, N., Jurisica, I.: Protein complex prediction via cost-based clustering. Bioinformatics **20**(17), 3013–3020 (2004)
172. King, R.D., Karwath, A., Clare, A., Dehaspe, L.: Accurate prediction of protein functional class from sequence in the mycobacterium tuberculosis and escherichia coli genomes using data mining. Yeast **17**(4), 283–293 (2000)
173. King, R.D., Karwath, A., Clare, A., Dehaspe, L.: The utility of different representations of protein sequence for predicting functional class. Bioinformatics **17**(5), 445–454 (2001)
174. Kirilova, S., Carugo, O.: Progress in the PRIDE technique for rapidly comparing protein three-dimensional structures. BMC Research Notes **1**, 44 (2008)
175. Kissinel, E., Henrick, K.: Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallographica D Bio Crystallogr **60**(12.1), 2256–2268 (2004)
176. Kleywegt, G.J.: Use of noncrystallographic symmetry in protein structure refinement. Acta Crystallogr D. **52**(Pt. 4), 842–857 (1996)
177. Koehl, P.: Protein structure similarities. Curr. Opinion Struct. Biol. **11**, 348–353 (2001)

178. Kolesnikov, N., et al.: Arrayexpress update–simplifying data submissions. Nucleic Acids Res **43**(Database issue), D1113–D1116 (2015)
179. Kolesov, G., Mewes, H.W., Frishman, D.: Snapping up functionally related genes based on context information: a colinearity-free approach. J Mol Biol **311**(4), 639–656 (2001)
180. Kolesov, G., Mewes, H.W., Frishman, D.: Snapper: gene order predicts gene function. Bioinformatics **18**(7), 1017–1019 (2002)
181. Kolodny, R., Koehl, P., Guibas, L., Levitt, M.: Small libraries of protein fragments model native protein structures accurately. J. Mol. Biol. **323**, 297–307 (2002)
182. Kolodny, R., Koehl, P., Levitt, M.: Comprehensive evaluation of protein structure alignment methods: Scoring by geometric measures. J. Mol. Biol. **346**, 1173–1188 (2005)
183. Kondor, R.I., Lafferty, J.: Diffusion kernels on graphs and other discrete structures. In: Int Conf Mach Learn (ICML), pp. 315–322 (2002)
184. Koonin, E.V., Galperin, M.Y.: Sequence - evolution - function: Computational approaches in comparative genomics. In: Evolutionary Concept in Genetics and Genomics, 1 edn., chap. 2. Kluwer Academic, Boston, MA (2003)
185. Korbel, J.O., Jensen, L.J., von Mering, C., Bork, P.: Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nature Biotechnol **22**(7), 911–917 (2004)
186. Koskinen, P., Törönen, P., Nokso-Koivisto, J., Holm, L.: PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. Bioinformatics **31**(10), 1544–1552 (2015)
187. Kourmpetis, Y.A., van Dijk, A.D., Bink, M.C., van Ham, R.C., ter Braak, C.J.: Bayesian markov random field analysis for protein function prediction based on network data. PLoS One **5**(2), e9293 (2010)
188. Kourmpetis, Y.A., van Dijk, A.D., ter Braak, C.J.: Gene ontology consistent protein function prediction: the falcon algorithm applied to six eukaryotic genomes. Algorithms Mol Biol **8**(1), 10 (2013)
189. Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., Leslie, C.: Profile-based string kernels for remote homology detection and motif extraction. J Bioinf Comput Biol **3**(3), 527–550 (2005)
190. Kuncheva, L.I., Bezdek, J.C., Duin, R.P.W.: Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction. Pattern Recognition **34**(2), 299–314 (2011)
191. Kunik, V., Solan, Z., Edelman, S., Ruppin, E., Horn, D.: Motif extraction and protein classification. In: Pacific Symp Biocomput (PSB), pp. 80–85 (2005)
192. Kuramochi, M., Karypis, G.: Gene classification using expression profiles. In: IEEE Symp Bioinf Bioeng (BIBE), pp. 191–200 (2001)
193. Lagreid, A., Hvidsten, T.R., Midelfart, H., Komorowski, J., Sandvik, A.K.: Predicting gene ontology biological process from temporal gene expression patterns. Genome Res **13**(5), 965–979 (2003)
194. Lan, L., et al.: Ms-knn: Protein function prediction by integrating multiple data sources. BMC Bioinform **14**(Suppl 1), S8 (2013)
195. Lanckriet, G.R.G., De Bie, T., Cristianini, N., Jordan, M.I., Noble, W.S.: A statistical framework for genomic data fusion. Bioinformatics **20**(16), 2626–2635 (2004)
196. Lanckriet, G.R.G., Deng, M., Cristianini, N., Jordan, M.I., Noble, W.S.: Kernel-based data fusion and its application to protein function prediction in yeast. In: Pacific Symp Biocomput (PSB), pp. 300–311 (2004)
197. Lavezzo, E., Falda, M., Fontana, P., Bianco, L., Toppo, S.: Enhancing protein function prediction with taxonomic constraints - the Argot2.5 web server. Methods **93**, 15–23 (2016)
198. Lee, D., Redfern, O., Orengo, C.: Predicting protein function from sequence and structure. Nat. Rev. Mol. Cell Biol. **8**, 995–1005 (2007)
199. Lee, J., Gross, S.P., Lee, J.: Improved network community structure improves function prediction. Scientific Reports **3**, 2197 (2013)
200. Lee, J., Lee, I., Lee, J.: Unbiased global optimization of Lennard-Jones clusters for n $\leq$ 201 using the conformational space annealing method. Phys Rev Lett **91**(8), 080,201 (2003)

201. Lee, J., Scheraga, H.A., Rackovsky, S.: New optimization method for conformational energy calculations on polypeptides: conformational space annealing. J Comput Chem **18**(9), 1222–1232 (1997)
202. Legrain, P., Wojcik, J., Gauthier, J.M.: Protein–protein interaction maps: a lead towards cellular functions. Trends Genet **17**(6), 346–352 (2001)
203. Leslie, C.S., Eskin, E., Cohen, A., Weston, J., Noble, W.S.: Mismatch string kernels for discriminative protein classification. Bioinformatics **20**(4), 467–476 (2003)
204. Letovsky, S., Kasif, S.: Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics **19**(Suppl 1), i197–i204 (2003)
205. Letsche, T.A., Berry, M.W.: Large-scale information retrieval with latent semantic indexing. Inf Sci **100**(1–4), 105–137 (1997)
206. Levitt, M., Gerstein, M.: A unified statistical framework for sequence comparison and structure comparison. Proc. Natl. Acad. Sci. USA **95**(11), 5913–5920 (1998)
207. Levy, E., Ouzounis, C.A., Gilks, W.R., Audit, B.: Probabilistic annotation of protein sequences based on functional classifications. BMC Bioinf **6**, 302 (2005)
208. Li, H., Liang, S.: Local network topology in human protein interaction data predicts functional association. PLoS One **4**(7), e6410 (2009)
209. Li, H., Tong, P., Gallegos, J., Dimmer, E., Cai, G., Molldrem, J.J., Liang, S.: PAND: A distribution to identify functional linkage from networks with preferential attachment property. PLoS One **10**(7), e0127,968 (15)
210. Li, H.L., Fujimoto, N., Sasakawa, N., Shirai, S., Ohkame, T., Sakuma, T., Tanaka, M., Amano, N., Watanabe, A., Sakurai, H., Yamamoto, T., Yamanaka, S., Hotta, A.: Precise correction of the dystrophin gene in duchenne muscular dystrophy patient induced pluripotent stem cells by TALEN and CRISPR-Cas9. Stem Cell Reports **4**(1), 143–154 (2015)
211. Li, L., Stoeckert, C.J., Roos, D.S.: OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res **13**(9), 2178–2189 (2003)
212. Li, Y., L., C.: Big biologica data: Challenges and opportunities. Genomics, Proteomics, and Bioinformatics **12**(5), 187–189 (2014)
213. Liao, L., Noble, W.S.: Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. J. Comp. Biol. **10**(6), 857–868 (2002)
214. Liberles, D.A., Thorn, A., von Heijne G. AN Elofsson, A.: The use of phylogenetic profiles for gene predictions. Current Genomics **3**(3), 131–137 (2002)
215. Lingling, A., Doerge, R.W.: Dynamic clustering of gene expression. ISRN Bioinformatics **2012**(537217), 1–12 (2012)
216. Lisewski, A.M., Lichtarge, O.: Rapid detection of similarity in protein structure and function through contact metric distances. Nucl. Acids Res. **34**(22), e152 (2006)
217. Liu, A.H., Califano, A.: Functional classification of proteins by pattern discovery and top-down clustering of primary sequences. IBM Systems J **40**(2), 379–393 (2001)
218. Liu, B., Wang, X., Chen, Q., Dong, Q., Lan, X.: Using amino acid physicochemical distance transformation for fast protein remote homology detection. PLoS One **7**(9), e46,633 (2012)
219. Liu, B., Wang, X., Lin, L., Dong, Q., Wang, X.: A discriminative method for protein remote homology detection and fold recognition combining top-n-grams and latent semantic analysis. BMC Bioinf **9**(510) (2008)
220. Liu, B., et al.: Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics **30**(4), 472–479 (2014)
221. Liu, J., Wang, W., Yang, J.: Gene ontology friendly biclustering of expression profiles. In: IEEE Comput Sys Bioinf Conf (CSB), pp. 436–447 (2004)
222. Liu, Q., Chen, Y.P., Li, J.: k-partite cliques of protein interactions: A novel subgraph topology for functional coherence analysis on PPI networks. J Theoretical Biol **340**(7), 146–154 (2014)
223. Lobley, A., Swindells, M.B., Orengo, C.A., Jones, D.T.: Inferring function using patterns of native disorder in proteins. PLoS Comput Biol **3**(8), e162 (2007)
224. Lobley, A.E.: Human protein function prediction: application of machine learning for integration of heterogeneous data sources. Ph.D. thesis, University College London (2010)

225. Lobley, A.E., Nugent, T., Orengo, C.A., Jones, D.T.: FFPred: an integrated feature-based function prediction server for vertebrate proteomes. Nucleic Acids Res **36**(Web server issue), W297–W302 (2008)
226. Ma, Q., Chirn, G.W., Cai, R., Szustakowski, J., Nirmala, N.C.: Clustering protein sequences with a novel metric transformed from sequence similarity scores and sequence alignments with neural networks. BMC Bioinf **6**(1), 242 (2005)
227. Ma, X., Chen, T., Sun, F.: Integrative approaches for predicting protein function and prioritizing genes for complex phenotypes using protein interaction networks. Briefings in Bioinformatics **15**(5), 685–698 (2013)
228. Maciag, K., et al.: Systems-level analyses identify extensive coupling among gene expression machines. Mol Syst Biol **2**(1), 0003 (2006)
229. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: A survey. IEEE Trans Comput Biol Bioinf **1**(1), 24–45 (2004)
230. Marchler-Bauer, A., et al.: CDD: a conserved domain database for protein classification. Nucleic Acids Res **33**(Database issue), D192–D196 (2005)
231. Marco, F., Alberto, B., Valentini, G.: UNIPred: Unbalance-aware network integration and prediction of protein functions. J Comput Biol **22**(12), 1057–1074 (2015)
232. Marcotte, C.J.V., Marcotte, E.M.: Predicting functional linkages from gene fusions with confidence. Applied Bioinf **1**(2), 93–100 (2002)
233. Marcotte, E.M., Pellegrini, M., Ng, H., Rice, D.W., Yeates, T.O., Eisenberg, D.: Detecting protein function and protein-protein interactions from genome sequences. Science **285**(5428), 751–753 (1999)
234. Marti-Renom, M.A., Capriotti, E., Shindyalov, I.N., Bourne, P.E.: Structure comparison and alignment. In: J. Gu, P.E. Bourne (eds.) Structural Bioinformatics, 2 edn., chap. 16. John Wiley & Sons (2009)
235. Martin, A.C.: The ups and downs of protein topology; rapid comparison of protein structure. Protein Eng. **13**(12), 829–837 (2000)
236. Martin, D.M., Berriman, M., Barton, G.J.: GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinf **5**(178) (2004)
237. Mateos, A., Dopazo, J., Jansen, R., Tu, Y., Gerstein, M., Stolovitzky, G.: Systematic learning of gene functional classes from dna array expression data by using multilayer perceptrons. Genome Res **12**(11), 1703–1715 (2002)
238. McDowall, M.D., Scott, M.S., Barton, G.J.: PIPs: human protein-protein interaction prediction database. Nucleic Acids Res **37**(Database issue), D651–D656 (2009)
239. Mi, H., Muruganujan, A., Casagrande, J.T., Thomas, P.T.: Large-scale gene function analysis with the PANTHER classification system. Nat Protocol **8**(8), 1551–1566 (2013)
240. Mi, H., et al.: The PANTHER database of protein families and subfamilies and functions and pathways. Nucleic Acids Res **33**(Database issue), D284–D288 (2005)
241. Midelfart, H., Laegreid, A., Komorowski, J.: Classification of gene expression data in an ontology. In: Medical Data Analysis, *Lecture Notes in Computer Science*, vol. 2199, pp. 186–194. Springer (2001)
242. Miele, V., Penel, S., Daubin, V., Picard, F., Kahn, D., Duret, L.: High-quality sequence clustering guided by network topology and multiple alignment likelihood. Bioinformatics **28**(8), 1078–1085 (2012)
243. Möller-Levet, C.S., Cho, K., Yin, H., Wolkenhauer, O.: Clustering of gene expression time-series data. Tech. rep., University of Rostock, Germany (2003)
244. Möller-Levett, C.S., Klawonn, F., Cho, K.: Clustering of unevenly sampled gene expression time-series data. Science **152**(1), 49–66 (2005)
245. Molloy, K., Min, J.V., Barbara, D., Shehu, A.: Exploring representations of protein structure for automated remote homology detection and mapping of protein structure space. BMC Bioinf **15**(Suppl 8), S4 (2014)
246. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn **52**(1), 91–118 (2003)

247. Moosavi, S., Rahgozar, M., Rahimi, A.: Protein function prediction using neighbor relativity in protein-protein interaction network. Comput Biol Chem **43**, 11–16 (2013)
248. Mostfavi, S., Morris, Q.: Fast integration of heterogeneous data sources for predicting gene function with limited annotation. Bioinformatics **26**(14), 1759–1765 (2010)
249. Muda, H.M., Saad, P., Othman, R.M.: Remote protein homology detection and fold recognition using two-layer support vector machine classifiers. Comput Biol Med **41**(8), 687–699 (2011)
250. Mukherjee, S.: Classifying microarray data using support vector machines. In: D.P. Berrar, W. Dubitzky, M. Granzow (eds.) A Practical Approach to Microarray Data Analysis, chap. 9. Kluwer Academic Publishers (2003)
251. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**, 536–540 (1995)
252. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M.: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics **21**(Suppl 1), i302–i310 (2005)
253. Nair, R., Carter, P., Rost, B.: Nlsdb: database of nuclear localization signals. Nucleic Acid Research **31**(1), 397–399 (2003)
254. Najmanovich, R.J., Torrance, W., Thornton, J.M.: Prediction of protein function from structure: Insights from methods for the detection of local structural similarities. Bio Techniques **38**(6), 847–851 (2005)
255. Nariai, N., Kolaczyk, E.D., Kasif, S.: Probabilistic protein function prediction from heterogeneous genome-wide data. PLoS One **2**(3), e337 (2007)
256. Narra, K., Liao, L.: Use of extended phylogenetic profiles with E-values and support vector machines for protein family classification. Intl J Computer Info Sci **6**(1) (2005)
257. Nepusz, T., Sasidharan, R., Paccanaro, A.: SCPS: a fast implementation of a spectral method for detecting protein families on a genome-wide scale. BMC Bioinf **11**(1), 120 (2010)
258. Ng, S., Tan, S., Sundararajan, V.: On combining multiple microarray studies for improved functional classification by whole-dataset feature selection. Genome Informatics **14**, 44–53 (2003)
259. Ng, S., Zhu, Z., Ong, Y.: Whole-genome functional classification of genes by latent semantic analysis on microarray data. In: Asia-Pacific Conf on Bioinformatics, pp. 123–129 (2004)
260. Ni, Q., Wang, Z., Han, Q., Li, G.: Using logistic regression method to predict protein function from protein-protein interaction data. In: IEEE Intl Conf Bioinf Biomed Eng (ICBBE), pp. 1–4 (2009)
261. Obozinski, G., Lanckriet, G., Grant, C., Jordan, M., Noble, W.S.: Consistent probabilistic output for protein function prediction. Genome Biol **9**(Suppl 1), S6 (2008)
262. Ofer, D., Linial, M.: ProFET: Feature engineering captures high-level protein functions. Bioinformatics **31**(21), 3429–3436 (2015)
263. Oliver, S.: Guilt-by-association goes global. Nature **403**(6770), 601–603 (2000)
264. Oliver, S.G.: From DNA sequence to biological function. Nature **379**(6566), 597–600 (1996)
265. Orchard, S., et al.: The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res **42**(Database issue), D358–D363 (2014)
266. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M.: CATH database: A hierarchic classification of protein domain structures. Structure **5**(8), 1093–1108 (1997)
267. Orengo, C.A., Taylor, W.R.: SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol **266**, 617–635 (1996)
268. Ortiz, A.R., Strauss, C.E., Olmea, O.: MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci **11**(11), 2606–2621 (2002)
269. Osadchy, M., Kolodny, R.: Maps of protein structure space reveal a fundamental relationship between protein structure and function. Proc. Natl. Acad. Sci. USA **108**, 12,301–12,306 (2011)

270. Ouali, M., King, R.D.: Cascaded multiple classifiers for secondary structure prediction. Protein Science **9**(6), 1162–1176 (2000)
271. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Matlsev, N.: Use of contiguity on the chromosome to predict functional coupling. In Silico Biol **1**(2), 93–108 (1999)
272. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Matlsev, N.: The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA **96**(6), 2896–2901 (1999)
273. Pagel, P., et al.: The MIPS mammalian protein-protein interaction database. Bioinformatics **21**(6), 832–834 (2005)
274. Pasquier, C., Promponas, V., Hamodrakas, S.J.: PRED-CLASS: cascading neural networks for generalized protein classification and genome-wide application. Proteins **44**(3), 361–369 (2000)
275. Pavlidis, P., Cai, J., Weston, J., Noble, W.S.: Learning gene functional classifications from multiple data types. J Comput Biol **9**(2), 401–411 (2002)
276. Pazos, F., Valencia, A.: Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng **14**(9), 609–614 (2001)
277. Pearl, F.M., Bennett, C.F., Bray, J.E., al., e.: The CATH database: an extended protein family resource for structural and functional genomics. Nucl. Acids Res. **31**, 452–455 (2003)
278. Pearson, W.R., Lipman, D.J.: Improved tools for biological sequence comparison. Proc Natl Aca Sci USA **85**(8), 2444–2448 (1988)
279. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O.: The underlying hypothesis is that two genes with similar phylogeny profiles will also be functionally similar. Proc Natl Acad Sci USA **96**(8), 4285–4288 (1999)
280. Pereira-Leal, J.B., Enright, A.J., Ouzounis, C.A.: Detection of functional modules from protein interaction networks. Proteins: Struct Funct Bioinf **54**(1), 49–57 (2004)
281. Pérez, A.J., Rodriguez, A., Trelles, O., Thode, G.: A computational strategy for protein function assignment which addresses the multidomain problem. Comp Funct Genomics **3**(5), 423–440 (2002)
282. Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, H., Will, G., North, A.C.T.: Structure of myoglobin: a three-dimensional fourier synthesis at 5.5 angstrom resolution. Nature **185**, 416–422 (1960)
283. Piovesan, D., Giollo, M., Ferrari, C., Tossato, S.C.E.: Protein function prediction using guilty by association from interaction networks. Amino Acids **47**(12), 2583–2592 (2015)
284. Prieto, C., De Las Rivas, J.: APID: Agile protein interaction dataanalyzer. Nucleic Acids Res **34**(Web Server issue), W298–W302 (2006)
285. Qian, B., Goldstein, R.A.: Detecting distant homologs using phylogenetic tree-based HMMs. Proteins **52**(3), 446–453 (2003)
286. Qin, W., Dion, S.L., Kutny, P.M., Zhang, Y., Cheng, A.W., Jillete, N.L., Malhotra, A., Geurts, A.M., Chen, Y.G., Wang, J.: Efficient CRISPR/Cas9-Mediated genome editing in mice by zygote electroporation of nuclease. Genetics **200**(2), 423–430 (2015)
287. Radivojac, P., et al.: A large-scale evaluation of computational protein function prediction methods. Nat Methods **10**(3), 221–227 (2013)
288. Rangwala, H., Karypis, G.: Profile-based direct kernels for remote homology detection and fold recognition. Bioinformatics **21**(23), 4239–4247 (2005)
289. Rappoport, N., Karsenty, S., Stern, A., Linial, N., Linial, M.P.: ProtoNet 6.0: organizing 10 million protein sequences in a compact hierarchical family tree. Nucleic Acids Res **40**(Database Issue), D313–D320 (2012)
290. Rawlings, N.D., Barrett, A.J.: MEROPS: the peptidase database. Nucleic Acids Res **27**(1), 325–331 (1999)
291. Raychaudari, S., Chang, J., Sutphin, P., Altman, R.: Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. Genome Research **12**(1), 203–214 (2002)
292. Re, M., Valentini, G.: Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction. J Mach Learn Res **8**, 98–111 (2010)

293. Remmert, M., Biegert, A., Hauser, A., Söding, J.: HHblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. Nat Methods **9**(2), 173–175 (2011)
294. Renner, A., Aszodi, A.: High-throughput functional annotation of novel gene products using document clustering. In: Proc. Symp. Biocomputing (PSB), pp. 54–68 (2000)
295. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. J Mach Learn **5**, 101–141 (2004)
296. Riley, M.: Systems for categorizing functions of gene products. Curr Opin Struct Biol **8**(3), 388–392 (1998)
297. Roch, K.G.L., et al.: Discovery of gene function by expression profiling of the malaria parasite life cycle. Science **301**(5639), 1503–1508 (2003)
298. Rogen, P., Fain, B.: Automatic classification of protein structure by using gauss integrals. Proc. Natl. Acad. Sci. USA **100**(1), 119–124 (2003)
299. Rost, B.: Enzyme function less conserved than anticipated. J Mol Biol **318**, 595–608 (1999)
300. Ruepp, A., et al.: The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res **32**(18), 5539–5545 (2004)
301. Saini, A., Hou, J.: Progressive clustering based method for protein function prediction. Bulletin Math Biol **75**(2), 331–350 (2013)
302. Samanta, M.P., Liang, S.: Predicting protein functions from redundancies in large-scale protein interaction networks. Proc Natl Acad Sci USA **100**(22), 12,579–12,583 (2003)
303. Sander, J.D., Joung, J.K.: CRISPR-Cas systems for editing, regulating and targeting genomes. Nature Biotechnology **32**(4), 347–355 (2014)
304. Sarac, O.S., Atalay, V., Cetin-Atalay, R.: GOPred: GO molecular function prediction by combined classifiers. PLoS One **5**(8), e12,382 (2010)
305. Sasson, O., Linial, N., Linial, M.P.: The metric space of proteins-comparative study of clustering algorithms. Bioinformatics **18**(Suppl 1), S14–S21 (2002)
306. Sboner, A., Mu, X.J., Greenbaum, D., Auerbach, R.K., Gerstein, M.B.: The real cost of sequencing: higher than you think! Genome Biol **12**(8), 125–134 (2011)
307. Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D., Dzeroski, S.: Predicting gene function using hierarchical multi-label decision tree ensembles. BMC Bioinf **11**(1), 2 (2010)
308. Schnoes, A.M., Brown, S.D., Dodevski, I., Babbitt, P.C.: Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol **5**(12), e1000,605 (2009)
309. Schnoes, A.M., Ream, D.C., Thorman, A.W., Babbitt, P.C., Friedberg, I.: Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. PLoS Comput Biol **9**(5), e1003,063 (2013)
310. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press (2002)
311. Schug, J.: Predicting gene ontology functions from ProDom and CDD protein domains. Genome Res **12**(4), 648–655 (2002)
312. Schwikowski, B., Uetz, P., Fields, S.: A network of protein-protein interactions in yeast. Nat Biotechnol **18**(12), 1257–1261 (2000)
313. Serres, M.H., Riley, M.: MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. Microb Comp Genomics **5**(4), 205–222 (2000)
314. Servant, F., Bru, C., Carrere, S., et al.: ProDom: Automated clustering of homologous domains. Briefings in Bioinformatics **3**(3), 246–251 (2002)
315. Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. Mol Sys Biol **3**(1), 88 (2007)
316. Sherlock, G., et al.: The stanford microarray database. Nucleic Acid Res **29**(1), 152–155 (2001)
317. Shi, X., et al.: BMRF-Net: a software tool for identification of protein interaction subnetworks by a bagging markov random field-based method. Bioinformatics **31**(14), 2412–2414 (2015)
318. Shiga, M., Takigawa, I., Mamitsuka, H.: Annotating gene function by combining expression data with a modular gene network. Bioinformatics **23**(13), i468–i478 (2007)

319. Shindyalov, I.N., Bourne, P.E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. **11**(9), 739–747 (1998)
320. Sierk, M.L., Pearson, W.R.: Sensitivity and selectivity in protein structure comparison. Protein Sci. **13**(3), 773–785 (2004)
321. Sjolanderk, K.: Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics **20**(2), 170–179 (2004)
322. Sliwoski, G., Kothiwale, S., Meiler, J., Lowe, E.W.: Computational method in drug discovery. Pharmacol Rev **66**(1), 334–395 (2014)
323. Soding, J.: Protein homology detection by HMM-HMM comparison. Bioinformatics **21**(7), 951–960 (2005)
324. Sokolov, A., Ben-Hur, A.: Hierarchical classification of gene ontology terms using the GOstruct method. J Bioinform Comput Biol **8**(2), 357–376 (2010)
325. Song, J., Singh, M.: How and when should interactome-derived clusters be used to predict functional modules and protein function? Bioinformatics **25**(23), 3143–3150 (2009)
326. Sonnenburg, S., Ratsch, G., Schafer, C., Scholkopf, B.: Large scale multiple kernel learning. journal of machine learning research. J Mach Learn Res **7**, 1531–1565 (2006)
327. Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., Durbin, R.: Pfam: Multiple sequence alignments and HMM-profiles of protein domains. Nucl. Acids Res. **26**(1), 320–322 (1998)
328. Sonnhammer, E.L., Eddy, S.R., Durbin, R.: Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins: Struct. Funct. Bioinf. **28**(3), 405–420 (1997)
329. Sonnhammer, E.L., Eddy, S.R., Durbin, R.: Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins **28**(3), 405–420 (1997)
330. Spirin, V., Mirny, L.A.: Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA **100**(21), 12,123–12,128 (2003)
331. Stark, A., Sunyaev, S., Russell, R.B.: A model for statistical significance of local similarities in structure. J. Mol. Biol. **326**(5), 1307–1316 (2003)
332. Subbiah, S., Laurents, D.V., Levitt, M.: Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Curr Biol **3**(3), 141–148 (1993)
333. Swift, S., Tucker, A., Vinciotti, V., Martin, N., Orengo, C., Liu, X., Kellam, P.: Consensus clustering and functional interpretation of gene-expression data. Genome Biol **5**(11), R94 (2004)
334. Szklarczyk, D., et al.: STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res **43**(Database Issue), D447–D552 (2015)
335. Tan, P., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Information Systems **29**, 293–313 (2004)
336. Tanay, A., Sharan, R., Kupiec, M., Shamir, R.: Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proc Natl Acad Sci USA **101**(9), 2981–2986 (2004)
337. Tang, L., Chen, J., Ye, J.: On multiple kernel learning with multiple labels. In: Intl Joint Conf Artif Intell (IJCAI), pp. 1255–1260 (2009)
338. Tang, M., et al.: Graphical models for protein function and structure prediction. In: M. Elloumi, A.Y. Zomaya (eds.) Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data, Wiley series on Bioinformatics: Computational Techniques nd Engineering, chap. 9, pp. 191–222. Wiley (2013)
339. Tarcea, V.G., et al.: Michigan molecular interactions r2: from interacting proteins to pathways. Nucleic Acids Res **37**(Database issue), D642–D646 (2009)
340. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., et al.: The COG database: an updated version includes eukaryotes. BMC Bioinf **4**, 41 (2003)
341. Tchagang, A.B., et al.: Mining biological information from 3D short time-series gene expression data: the OPTricluster algorithm. BMC Bioinf **13**(54), 2105–2154 (2012)
342. Tetko, I., Facius, A., Ruepp, A., Mewes, H.W.: Super paramagnetic clustering of protein sequences. BMC Bioinf **6**(1), 82 (2005)
343. Thode, G., Garcia-Ranea, J.A., Jimenez, J.: Search for ancient patterns in protein sequences. J Mol Evol **42**(2), 224–233 (1996)
344. Thomas, T.: Multidomain proteins. eLS pp. 1–8 (2014)

345. Thoren, A.: The PhylProm database - extending the use of phylogenetic profiles and their applications for membrane proteins. Master's thesis, Stockholm University, Sweden (2000)
346. Tordai, H., Nagy, A., Farkas, K., Bányai, L., Patthy, L.: Modules, multidomain proteins and organismic complexity. FEBS J **272**(19), 5064–5078 (2005)
347. Tornow, S., Mewes, H.W.: Functional modules by relating protein interaction networks and gene expression. Nucleic Acids Res **31**(21), 6283–6289 (2003)
348. Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., Botstein, D.: A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae. Proc Natl Acad Sci USA **100**(4), 8348–8353 (2003)
349. Tsai, C.J., Nussinov, R.: Hydrophobic folding units at protein-protein interfaces: implications to protein folding and to protein-protein association. Protein Sci **6**(7), 1426–1437 (1996)
350. Uchiyama, I.: Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. Nucleic Acids Res **34**(2), 647–658 (2006)
351. Valastyan, J.S., Lindquist, S.: Mechanisms of protein-folding diseases at a glance. Disease Models and Mechanisms **7**(1), 9–14 (2014)
352. Valentini, G.: True path hierarchical ensembles for genome-wide gene function prediction. IEEE Trans Comput Biol Bioinform **8**(3), 832–847 (2011)
353. van Noort, V., Snel, B., Huynen, M.A.: Predicting gene function by conserved co-expression. Trends Genet **19**(5), 238–242 (2003)
354. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., Sharan, R.: Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol **6**(1), e1000,641 (2010)
355. Vazquez, A., Flammini, A., Maritan, A., Vespignani, A.: Global protein function prediction from protein-protein interaction networks. Nature Biotechnol **21**(6), 697–700 (2003)
356. Veretnik, S., Gu, J., Wodak, S.: Identifying structural domains in proteins. In: J. Gu, P. Bourne (eds.) Structural Bioinformatics, 2 edn., chap. 20, pp. 487–515. John Wiley & Sons (2009)
357. Verleyen, W., Ballouz, S., Gillis, J.: Measuring the wisdom of the crowds in network-based gene function inference. Bioinformatics **31**(5), 745–752 (2015)
358. Vert, J.: A tree kernel to analyze phylogenetic profiles. Bioinformatics **18**(Suppl 1), S276–S284 (2002)
359. Vlahovicek, K., Murvai, J., Barta, E., Pongor, S.: The SBASE protein domain library and release 9.0: an online resource for protein domain identification. Nucleic Acids Res **30**(1), 273–275 (2002)
360. Vlahovicek, K., Pintar, A., Parthasarathi, L., Carugo, O., Pongor, S.: CX, DPX and PRIDE: WWW servers for the analysis and comparison of protein 3d structures. Nucleic Acids Res **33**(Web Server issue), W252–W254 (2005)
361. Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C., Teichmann, S.A.: Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol **14**(2), 208–216 (2004)
362. Walker, M.G., Volkmuth, W., Sprinzak, E., Hodgson, D., Klingler, T.: Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. Genome Res **9**(12), 1198–1203 (1999)
363. Wang, D., Hou, J.: Explore the hidden treasure in protein-protein interaction networks - an iterative model for predicting protein functions. J Bioinf and Comput Biol **13**(1550026), 22 (2015)
364. Wang, M., Shang, X., Xie, D., Li, Z.: Mining frequent dense subgraphs based on extending vertices from unbalanced PPI networks. In: IEEE Intl Conf Bioinf Biomed Eng (ICBBE), pp. 1–7 (2009)
365. Wang, X., Schroeder, D., Dobbs, D., Honavar, V.: Automated data-driven discovery of motif-based protein function classifiers. Inf Sci **155**(1–2), 1–18 (2003)
366. Wang, Z., Cao, R., Cheng, J.: Three-level prediction of protein function by combining profile-sequence search, profile-profile search, and domain co-occurrence networks. BMC Bioinf **14**(3), S3 (2013)
367. Wass, M.N., Barton, G., Sternberg, M.J.E.: Combfunc: predicting protein function using heterogeneous data sources. Nucleic Acids Res **40**(Web server issue), W466–W470 (2012)

368. Wass, M.N., Sternberg, M.J.: ConFunc-functional annotation in the twilight zone. Bioinformatics **24**(6), 798–806 (2007)
369. Whisstock, J.C., Lesk, A.M.: Prediction of protein function from protein sequence and structure. Q Rev Biophys **36**(3), 307–340 (2003)
370. Wohlers, I., Andonov, R., Klau, G.W.: Algorithm engineering for optimal alignment of protein structure distance matrices. Optimization Letters (2011). DOI 10.1007/s11590-011-0313-3. URL https://hal.inria.fr/inria-00586067
371. Wohlers, I., Le Boudic-Jamin, M., Djidjev, H., Klau, G.W., Andonov, R.: Exact Protein Structure Classification Using the Maximum Contact Map Overlap Metric. In: 1st International Conference on Algorithms for Computational Biology, AlCoB 2014, pp. 262–273. Tarragona, Spain (2014). DOI 10.1007/978-3-319-07953-0_21. URL https://hal.inria.fr/hal-01093803
372. Wohlers, I., Malod-Dognin, N., Andonov, R., Klau, G.W.: CSA: Comprehensive comparison of pairwise protein structure alignments. Nucleic Acids Research pp. 303–309 (2012). URL https://hal.inria.fr/hal-00667920. Preprint, submitted to Nucleic Acids Research
373. Wu, C., Berry, M., Shivakumar, S., McLarty, J.: Neural networks for full-scale protein sequence classification: Sequence encoding with singular value decomposition. Mach Learn **21**(1), 177–193 (1992)
374. Wu, C., Ermongkonchai, A., Chang, T.C.: Protein classification using a neural network proein database (nnpdb) system. In: Anal Neural Net Appl Conf, pp. 29–41 (1991)
375. Wu, C., Whitson, G., McLarty, J., Ermongkonchai, A., Chang, T.C.: Protein classification artificial neural system. Protein Sci **1**(5), 667–677 (1995)
376. Wu, C.H., Whitson, G.M., Montllor, G.J.: PROCANS: a protein classification system using a neural network. Neural Networks **2**, 91–96 (1990)
377. Wu, J., Kasif, S., DeLisi, C.: Identification of functional links between genes using phylogenetic profiles. Bioinformatics **19**(12), 1524–1530 (2003)
378. Wu, L.F., Hughes, T.R., Davierwala, A.P., Robinson, M.D., Stoughton, R., Altschuler, S.J.: Large-scale prediction of saccharomyces cerevisiae gene function using overlapping transcriptional clusters. Nat Genet **31**(3), 255–265 (2002)
379. Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., Eisenberg, D.: Dip: the database of interacting proteins. Nucleic Acids Res **28**(1), 289–291 (2000)
380. Xie, H., Wasserman, A., Levine, Z., Novik, A., Grebinskiy, V., Shoshan, A., Mintz, L.: Large-scale protein annotation through gene ontology. Genome Res **12**(5), 785–794 (2002)
381. Yahalom, R., Reshef, D., Wiener, A., Frankel, S., Kalisman, N., Lerner, B., Keasar, C.: Structure-based identification of catalytic residues. Proteins **79**(6), 1952–1963 (2011)
382. Yan, Y., J., M.: Protein family clustering for structural genomics. J Mol Biol **353**(3), 744–759 (2005)
383. Yanai, I., Derti, A., DeLisi, C.: Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc Natl Acad Sci USA **98**(14), 7940–7945 (2001)
384. Yang, J., Wang, H., Wang, W., Yu, P.: Enhanced biclustering on expression data. In: IEEE Symp Bioinf Bioeng (BIBE), pp. 321–327 (2003)
385. Yona, G., Linial, N., Linial, M.P.: ProtoMap: automatic classification of protein sequences and hierarchy of protein families. Nucleic Acids Res **28**(1), 49–55 (2000)
386. Yu, G., Rangwala, H., Domeniconi, C., Zhang, G., Yu, Z.: Protein function prediction using multi-label ensemble classification. IEEE/ACM Trans Comput Biol Bioinform **10**(4), 1045–1057 (2013)
387. Zemla, A.: LGA: a method for finding 3D similarities in protein structures. Nucl. Acids Res. **31**(13), 3370–3374 (2003)
388. Zhang, W., et al.: The functional landscape of mouse gene expression. J Biol **3**(5), 21 (2004)
389. Zhang, X., Dai, D.: A framework for incorporating functional interrelationships into protein function prediction algorithms. IEEE/ACM Trans Comput Biol Bioinform **9**(3), 740–753 (2012)
390. Zhang, Y., Skolnick, J.: TM-align: a protein structure alignment algorithm based on the TM-score. Nucl. Acids Res. **33**(7), 2302–2309 (2005)

391. Zhang, Z.H., Hwee, K.L., Mihalek, I.: Reduced representation of protein structure: implications on efficiency and scope of detection of structural similarity. BMC Bioinformatics **11**, 155 (2010)
392. Zheng, Y., Roberts, R.J., Kasif, S.: Genomic functional annotation using co-evolution profiles of gene clusters. Genome Biol **3**(11), research0060.1–0060.9 (2002)
393. Zhou, D., Bousquet, O., Lal, T., Weston, J., Schlkopf, B.: Learning with local and global consistency. In: Advances Neural Inform Processing Systems (NIPS), pp. 321–328 (2004)
394. Zhou, X., Kao, M.C., Wong, W.: Transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci USA **99**(20), 12,783–12,788 (2002)
395. Zhou, Y., Young, J.A., Santrosyan, A., Chen, K., Yan, S.F., Winzeler, E.A.: In silico gene function prediction using ontology-based pattern identification. Bioinformatics **21**(7), 1237–1245 (2005)
396. Zhu, J., Zhang, M.Q.: SCPD: a promoter database of the yeast saccharomyces cerevisiae. Bionformatics **15**(7), 607–611 (1999)
397. Zitnik, M., Zupan, B.: Data fusion by matrix factorization. IEEE Trans Pattern Anal Mach Intell **37**(1), 41–53 (2015)