frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 29 September 2022
DOI 10.3389/fgene.2022.969915

Check for updates

OPEN ACCESS

EDITED BY
Jiyao Wang,
National Center for Biotechnology Information, National Library of Medicine (NIH), United States

REVIEWED BY
Min Zeng,
Central South University, China
Wenzheng Bao,
Xuzhou University of Technology, China

*CORRESPONDENCE
Subhadip Basu,
subhadip.basu@jadavpuruniversity.in
Dariusz Plewczynski,
dariuszplewczynski@cent.uw.edu.pl

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 15 June 2022
ACCEPTED 31 August 2022
PUBLISHED 29 September 2022

CITATION
Sengupta K, Saha S, Halder AK, Chatterjee P, Nasipuri M, Basu S and Plewczynski D (2022), PFP-GO: Integrating protein sequence, domain and protein-protein interaction information for protein function prediction using ranked GO terms. *Front. Genet.* 13:969915.
doi: 10.3389/fgene.2022.969915

PFP-GO: Integrating protein sequence, domain and protein-protein interaction information for protein function prediction using ranked GO terms

Kaustav Sengupta[1,2,3†], Sovan Saha[4†], Anup Kumar Halder[1,3], Piyali Chatterjee[5], Mita Nasipuri[2], Subhadip Basu[2*] and Dariusz Plewczynski[1,3*]

[1]Laboratory of Functional and Structural Genomics, Center of New Technologies, University of Warsaw, Warsaw, Poland, [2]Department of Computer Science and Engineering, Jadavpur University, Kolkata, India, [3]Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland, [4]Department of Computer Science and Engineering, Institute of Engineering and Management, Kolkata, West Bengal, India, [5]Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India

Protein function prediction is gradually emerging as an essential field in biological and computational studies. Though the latter has clinched a significant footprint, it has been observed that the application of computational information gathered from multiple sources has more significant influence than the one derived from a single source. Considering this fact, a methodology, PFP-GO, is proposed where heterogeneous sources like Protein Sequence, Protein Domain, and Protein-Protein Interaction Network have been processed separately for ranking each individual functional GO term. Based on this ranking, GO terms are propagated to the target proteins. While Protein sequence enriches the sequence-based information, Protein Domain and Protein-Protein Interaction Networks embed structural/functional and topological based information, respectively, during the phase of GO ranking. Performance analysis of PFP-GO is also based on Precision, Recall, and F-Score. The same was found to perform reasonably better when compared to the other existing state-of-art. PFP-GO has achieved an overall Precision, Recall, and F-Score of 0.67, 0.58, and 0.62, respectively. Furthermore, we check some of the top-ranked GO terms predicted by PFP-GO through multilayer network propagation that affect the 3D structure of the genome. The complete source code of PFP-GO is freely available at https://sites.google.com/view/pfp-go/.

KEYWORDS

protein sequence, protein domain, protein-protein interaction network, 3D gene-gene association, ranked GO, protein function prediction

Introduction

In recent years, protein function prediction has started using integrated function predictive information from several sources (Piovesan et al., 2015) instead of using a single source of information. These include Protein-Protein Interaction (PPI), Protein domain, Amino Acid Sequence of protein, Protein's structure, Genomic information, etc. Though integrated information enriched classifiers should perform better than a single type of feature, the design of such a single classification system of heterogeneous data sources is challenging for the Proteomic research community. Moreover, the prediction task becomes more challenging as it is characterized by several factors: 1) Any protein may be associated with multiple functions, i.e., one-to-many relationships, 2) functional groups are numerous, and 3) their existence is hierarchically structured and unbalanced.

The functionality of a protein can be attributed to the physical interactions represented in Protein-Protein Interactions (PPIs). Using proximity relationships between connected proteins, computational methods attempt to propagate the labels of known proteins to unknown proteins across the network. PPIs are mediated by their constituent domain-domain interactions (Chatterjee et al., 2011). Genes evolved from the same ancestral gene are functionally similar. So, finding known genes with sufficient sequence similarity is a powerful way to predict function. These various types of data individually are not sufficient to annotate functional groups.

Moreover, a recent trend shows that hierarchical relationships between functional classes motivate the development of hierarchy-aware prediction methods, which are significantly better than hierarchy-unaware "flat" prediction methods (Pandey et al., 2006). Functional relationships, e.g., Taxonomy like Gene Ontology (GO) (Ashburner et al., 2006) or FunCat (Ruepp et al., 2004), can be exploited to improve the predictive performance of learning algorithms.

Motivated by the facts mentioned above, here we propose an integrated approach where their orthogonal methods, namely, constituent domains of the protein and their interactions, sequence homology, and protein interaction data, are used to assign functional GO terms for unknown proteins and then these prediction decisions are combined into a consensus decision using n-star approach and functional enrichment. The following section discusses the current state of the arts of computational techniques in the protein function prediction domain based on raw amino sequence, domain, and protein interaction data.

Sequence-based approaches for protein function prediction

Genes that evolved through duplication and rearrangement from single ancestral genes are known to be homologous. The homologous genes are found at various places in the same genome. However, duplication is considered paralogous, whereas some orthologous genes diverge through speciation events found in different organisms. Inference of functional terms from sequence similarity is well supported by Anfinsen's dogma claiming protein's sequence determines protein's tertiary structure (Anfinsen, 1973). The basic strategy of the sequence-based prediction method is that similar known proteins are searched from a database for any target protein, assigning associated GO terms to that protein of interest. Local alignment-based tools like SSEARCH (Pearson, 1995) take a target sequence and find top hit sequences along with their statistically significant score E-value with the Smith-Waterman dynamic programming algorithm. As it is time-consuming, FASTA (Pearson and Lipman, 1988) does pairwise alignments only on highly similar regions using a lookup table, and BLAST (Altschul et al., 1990) saves time with the use of pre-computed similar words. However, these strategies are not sensitive to all protein families with different conservation degrees. On the other hand, PSI-BLAST (Altschul et al., 1997) uses sequence profile instead of raw sequence. It makes a profile of the target sequence and similar sequence at each iteration and uses a computed profile at the next iteration.

Pre-computed profiles of protein domains or portions of the conserved region can be used for assigned tasks. BLOCKS (Pietrokovski et al., 1996), ProDom (Corpet et al., 2000), Pfam (Finn et al., 2016), and SUPERFAMILY (Pandit et al., 2002) are datasets of profiles of protein domains, PRINTS (Attwood, 2002) is collection of protein fingerprints. Here target sequence and similar sequences are represented in the profile where the target protein profile is matched against the database sequence profile.

Sequence-based prediction is easy to use because most proteins are available with their sequence and functional annotation. However, limitations to sequence-based methods arise when 60% of the sequences are similar (Kihara, 2011). A correlation between structural and functional similarity can be used in that case. Sequence-based prediction methods are helpful when only raw protein sequence information is available. However, sometimes it becomes challenging because wrong functional annotation may be propagated for functional assignment, or correct function prediction does not always take place as important issues are not always considered, like non-orthologous displacement of genes or proteins has multiple domains (Chitale et al., 2009; Halder et al., 2019).

Domain-based approaches for protein function prediction

Protein domains are independent units of protein function which have unique three-dimensional structures. Protein functions are collective results of functions of its constituent domains. So, to predict function, exploiting the domain architecture of proteins is the need of the hour and their interaction and cooperation pattern. Some current *state-of-the-art* techniques use domain information for function prediction. Peng et al. (2014) use protein domain information along with Protein interaction networks and complexes. The domain combination similarity (DCS) representing the domain compositions of both proteins and their neighbors is used in their algorithm. Robert Rentzsch and Christine A Orengo (Rentzsch and Orengo, 2013) derive a functional family by combining sequence clustering with supervised cluster evaluation.

INGA (Piovesan et al., 2015) is another predictor using domain architecture and transfer annotation from proteins sharing the same domain pattern. It identifies putative PFAM domains, and all proteins associated with GO terms from UniProt are retrieved, and finally, those GO terms are assigned to the target protein (Piovesan et al., 2015).

Protein interaction network-based approaches for prediction of protein function

Protein interactions have great importance in protein function, so the function of an unknown protein can be extrapolated from the functional annotation of its interaction profiles exploiting proximity relationships. There are mainly four categories of network-based approaches in functional inferences of protein: Markov random field based, optimization-based, Clustering based, and neighbor based (Pandey et al., 2006). The trend also combines decisions about protein functions obtained through different approaches. Neighborhood-based approaches are primarily based on the idea that the proteins closer to the network are more similar in functionality. In prior studies (Schwikowski et al., 2000; Hishigaki et al., 2001), the functionality of target proteins is assigned considering the probability of occurrence of functions among the neighboring proteins. However, these approaches limited the neighboring proteins to level-1 of the target proteins in assigning the functional annotations. In another study (Chen et al., 2007) of protein function, further advancement has been introduced in the neighbor-based approach by introducing the network motifs concept in protein interactome. A global optimization mechanism was introduced in functional assignments to its unclassified (target) partners in the PPIN (Vazquez et al., 2003). In other work (Karaoz et al., 2004), the functional linkage graphs have been mapped into a variant of a discrete-state Hopfield network in order to gain the maximally consistent assignment by minimizing an "energy" function that includes a heuristic-guided local search mechanism. However, Karaoz et al. (2004) focused more on the global properties of interaction maps but not on the local proximity of interacting proteins (Nabieva et al., 2005). To overcome these above-described issues and considering the distant effects of annotated proteins, Nabieva et al. (2005) have introduced a Functional Flow based strategy using network flow where each protein of known function is annotated as a "source" of "functional flow." The work of Deng, Mehta et al. (2002) and Letovsky and Kasif (2003) are worth mentioning among probabilistic approaches. Functional module detection (Bader and Hogue, 2003; Sharan et al., 2007) and graph clustering methods (Spirin and Mirny, 2003; King et al., 2004) are effective module-assisted approaches.

Integrated approaches based on sequence, domain, and protein-protein interaction networks for protein function prediction

Inference of functional terms from sequence similarity is well supported by Anfinsen's dogma claiming protein's sequence determines protein's tertiary structure (Anfinsen, 1973). The basic strategy of the sequence-based prediction method is that for any target protein, similar known proteins are searched from a database using sequence similarity tools, like, PSI-BLAST, and assign those associated GO terms to that protein of interest. Local alignment-like tools take top hit sequences with a significant E-value or Smith-Waterman alignment score. A global or local alignment algorithm is used to infer homology (Altschul et al., 1997). DNA binding site prediction is also considered to be sequence-based function prediction. Sequence information and various types of sequence-derived features, Physico-chemical properties (for example, polarity, hydrophobicity), PSSM, the composition of amino acid, dipeptide composition, structural features like secondary structure, solvent accessible surface area are used (Altschul et al., 1997; Halder et al., 2019; Pearson, 1995; Pearson & Sierk, 2005). Protein subcellular location prediction (Garg et al., 2005; Sarda et al., 2005), enzyme function prediction (Wang et al., 2010), and signal peptide prediction (Nielsen et al., 1997) can also facilitate the prediction of protein function. The use of machine learning algorithms (Wang and Xiao, 2014; Xiao et al., 2012) and nearest neighbor classifier (Huang and Li, 2004; Wang and Yang, 2010) is significant in this regard. Sequence-based prediction is easy to use because most proteins are available with their sequence and functional annotation. However, limitations to sequence-based methods arise if two sequences have similarities below 60% (Kihara, 2011). A correlation between structural and functional similarity can be used in that case.

TABLE 1 Current computational methodologies of protein function prediction.

Features used	Brief description	References
Sequence and Network	A deep learning framework for gene ontology annotations with sequence- and network-based information	F. Zhang et al., (2020)
	DeepFunc: A deep learning framework for accurate prediction of protein functions from protein sequences and interactions	F. Zhang et al., (2019)
	Predicting GO annotations from protein sequences and interactions	X. Zhang et al., (2021)
GO terms	A deep learning framework for predicting protein functions with co-occurrence of GO terms	M. Li et al., (2022)
	Gene function prediction based on gene ontology hierarchy preserving hashing	Zhao et al. (2019)
	Gene function prediction based on combining gene ontology hierarchy with multi-instance multi-label learning	Z. Li et al., (2018)
Structure	Structure-based protein function prediction using graph convolutional networks	Gligorijević et al. (2021)
	Structure-based function prediction: approaches and applications	Gherardini and Helmer-Citterich, (2008)
	prediction of protein function from structure: insights from methods for the detection of local structural similarities	Najmanovich et al. (2005)

Prediction of protein function can also be performed by domain information of protein. Protein domains are independent folding units that represent basic functional units of protein. In another work (Rentzsch and Orengo, 2013), the function prediction is made using domain families derived through sequence clustering. Forslund and Sonnhammer (2008) computationally assign GO terms to unknown proteins based on the presence of identifiable domains. Rule-based and probabilistic models investigate the dependence between protein domain content and function.

Proteins perform their functions through interaction with each other. Moreover, a protein is associated with multiple functions. So, inference of function for any unknown protein can be made from its interaction information with its interacting partner. Recently, computational function prediction techniques are gaining importance from PPIN also. Network-based approaches are classified into the following groups: neighbor-based (Saha et al., 2014), optimization-based (Chen and Xu, 2004; Deng, Sun, et al., 2002) (Schwikowski et al., 2000), Markov random field based (Deng, Sun, et al., 2002) and Clustering based (Dandekar et al., 1998).

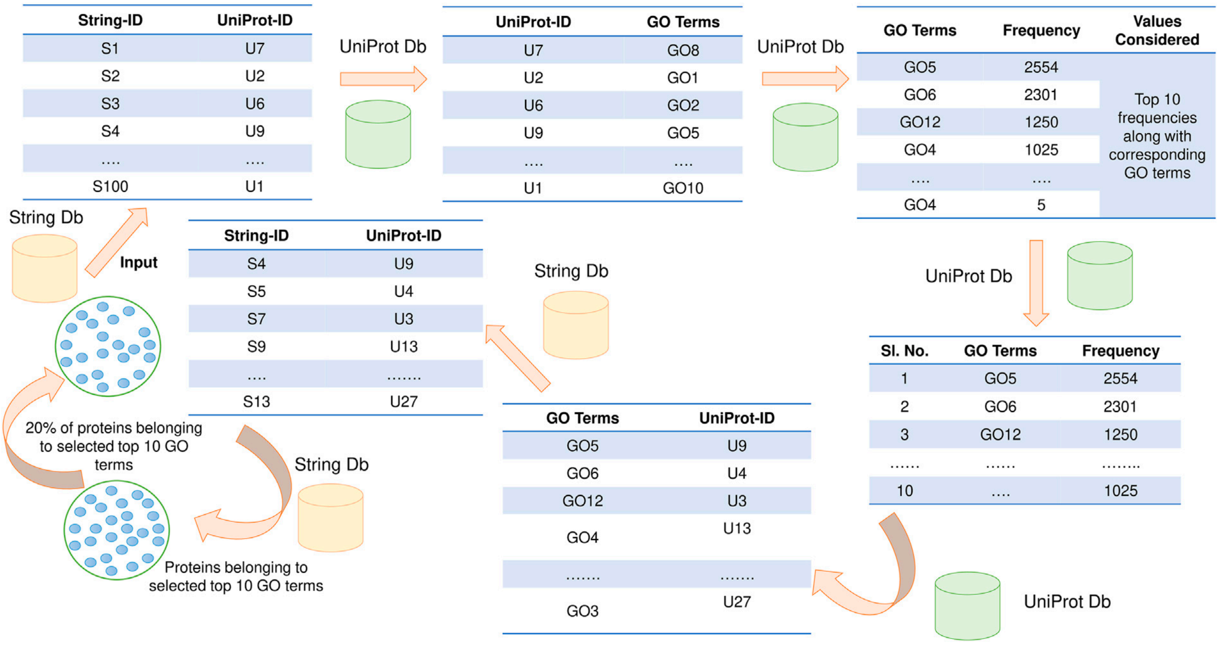
Some of the other relevant works of protein function prediction have been summarized in Table 1. Besides, some advanced studies use multi-features obtained from protein and sequence (Bao et al., 2017, Bao et al., 2021, Bao et al., 2022). Considering all these existing works, it is observed that there is still a pressing need to explore specific areas where heterogeneous sources are blended for a common cause of predicting protein function. This work proposes a methodology named PFP-GO to capture the protein functional dependencies on the domain, sequence, and PPI. The main idea is to embed all available information sources to consider other essential features rather than using sequence, domain, or PPI alone. All the source code of PFP-GO is available on https://sites.google.com/view/pfp-go/.

Methodology

PFP-GO assigns functional groups to target proteins based on the information integration of sequence similarity, PPI networks, and domain assignments. This method combines these orthogonal predictions and derives consensus predictions for GO terms using functional enrichment. As different heterogeneous information sources are used, mapping data from one source to another is essential in this regard. Protein interaction networks may contain specific false positive and false negative data. So, finding biologically essential proteins is challenging as they are promising candidates for finding drug targets. This work categorizes PFP-GO into four sections: 1) It identifies the functionally active target proteins (i.e., proteins associated with frequently occurring GO terms) whose functional groups are predicted. 2) Each target protein's level-2 neighborhood graph is considered, and non-essential proteins, i.e., shore, bridge, and fjord proteins (Hanna and Zaki, 2014), are eradicated. 3) Once the refined PPI for each target protein is obtained, sequence-based, domain-based, and neighborhood protein interaction-based approaches are applied to the target and its neighbors to assign GO terms. 4) GO terms are ranked using a functional enrichment score. 5) Finally, common GO terms among the sequence-based, domain-based, and neighborhood protein interaction-based prediction results are finally transmitted to the target protein following a 3-star consensus (Chatterjee et al., 2016) approach.

Database

PFP-GO can be centrally categorized into three sections: 1) Sequence-based prediction, 2) Domain-Domain interaction-based prediction, and 3) Topology or neighborhood-based prediction from the PPI network. In topology or



String-ID	UniProt-ID
S1	U7
S2	U2
S3	U6
S4	U9
....	....
S100	U1

UniProt-ID	GO Terms
U7	GO8
U2	GO1
U6	GO2
U9	GO5
....	....
U1	GO10

GO Terms	Frequency	Values Considered
GO5	2554	Top 10 frequencies along with corresponding GO terms
GO6	2301	
GO12	1250	
GO4	1025	
....	....	
GO4	5	

Sl. No.	GO Terms	Frequency
1	GO5	2554
2	GO6	2301
3	GO12	1250
......	......	....
10	....	1025

GO Terms	UniProt-ID
GO5	U9
GO6	U4
GO12	U3
GO4	U13
....	....
GO3	U27

String-ID	UniProt-ID
S4	U9
S5	U4
S7	U3
S9	U13
....	....
S13	U27

UniProt Db
UniProt Db
String Db
Input
String Db
UniProt Db
UniProt Db
String Db
20% of proteins belonging to selected top 10 GO terms
Proteins belonging to selected top 10 GO terms

FIGURE 1
Functionally active target protein selection. Two databases: STRING and UniProt, have been used for this purpose.

neighborhood-based prediction from PPI networks, String (Franceschini et al., 2013) and Uniprot (Consortium, 2015) databases are used to generate PPI network and GO terms, respectively. In domain-domain interaction-based prediction, PFAM (Finn et al., 2016) and DOMINE database (Yellaboina et al., 2011) are used for obtaining the domain-domain interaction from the corresponding Uniprot ids. GO Consortium database (Consortium, 2018) also plays an essential role in this section in including its own GO terms. While in sequence-based prediction, STRING and Uniprot are utilized in coordination for protein sequence and GO term generation.

Identification of functionally active targets

In this section, the STRING database id used to fetch interactions (Franceschini et al., 2013), and the UniProt database was used for GO annotations (U. Consortium, 2015). If direct mapping from STRING ID to UniProt ID is unavailable, then PFP-GO focuses on identifying a homogenous entry with at least 90% sequence identity from UniProt. Next, computation of the frequency of associated GO terms for every protein is implemented. The top 10 frequently occurring GO terms are considered based on the frequency of GO terms. The STRING IDs are fetched from these corresponding GO terms using UniProt as an intermediary. 20% of these proteins (STRING IDs) are randomly considered to be functionally active target proteins. The schematic diagram of selecting active target proteins is highlighted in Figure 1.

Pruning and filtering of target protein neighborhood graph

For each target protein, interaction information is retrieved from the STRING database (Franceschini et al., 2013), and their neighborhood graph (consisting of level-1 and level-2 proteins) is formed. To remove non-essential protein, a topology-based method is considered for testing the target protein neighbor's essentiality. In order to assess the essentiality, it is checked whether any protein in the target's neighborhood is of bridge or Fjord or shore protein (Hanna and Zaki, 2014). These neighbors get ultimately pruned to ensure that their presence might not affect the prediction accuracy (see Figure 2).

Besides this bridge, fjord and shore proteins, Network centrality-based Edge Clustering Coefficient (ECC) (Peng et al., 2012), and edge-weight (S. Wang and Wu, 2013) are yet another two most effective measures for the identification of essential proteins. ECC measures the degree of closeness between two nodes in a graph. Those edges with higher ECC values are more likely to be in a module. In comparison, edge-weight assigns a weightage to each edge, which signifies the reliability of the corresponding edge. So double filtering using both ECC

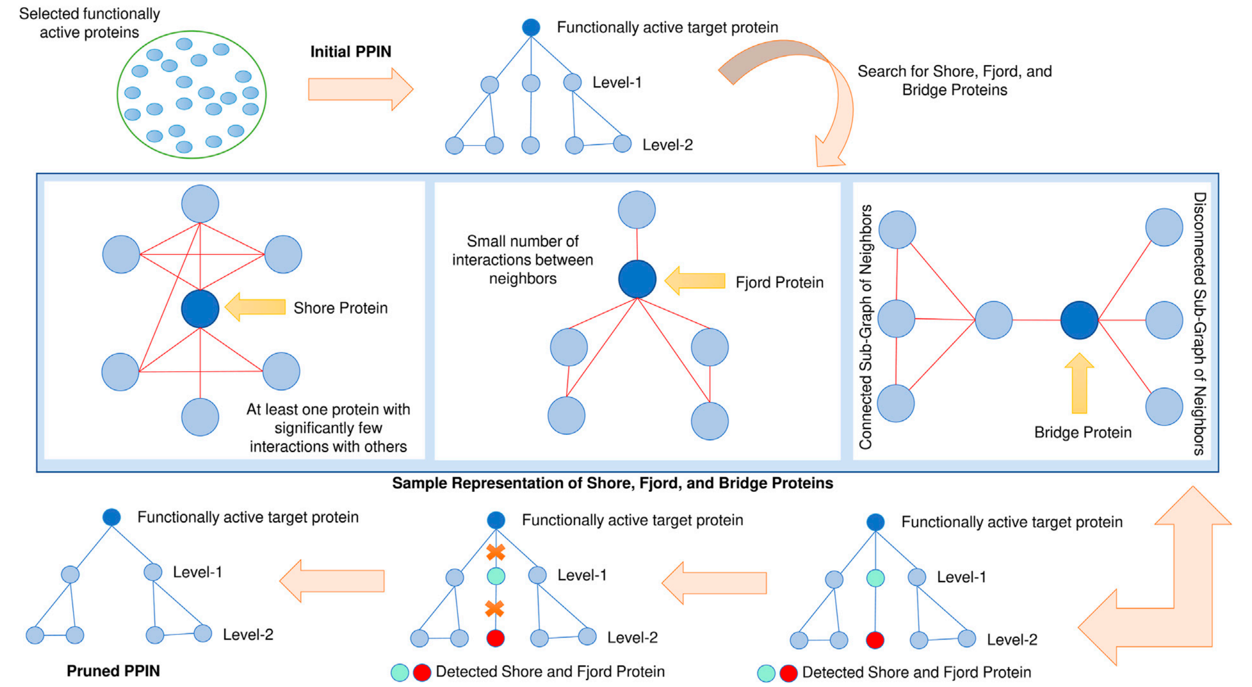

Selected functionally active proteins
Initial PPIN
Functionally active target protein
Level-1
Level-2
Search for Shore, Fjord, and Bridge Proteins
Shore Protein
At least one protein with significantly few interactions with others
Small number of interactions between neighbors
Fjord Protein
Connected Sub-Graph of Neighbors
Disconnected Sub-Graph of Neighbors
Bridge Protein
Sample Representation of Shore, Fjord, and Bridge Proteins
Functionally active target protein
Level-1
Level-2
Pruned PPIN
Detected Shore and Fjord Protein
Detected Shore and Fjord Protein

FIGURE 2
Pruning of target neighborhood graphs. Bridge, fjord, and shore proteins are detected and pruned.

and edge-weight is also implemented on the pruned target neighborhood network to ensure the presence of the most reliable edges. The schematic diagram of this 2-pass filtering approach has been highlighted in Figure 3. The threshold of both ECC and edge-weight is calculated by the following mathematical equation (Zhang et al., 2016):

$$Threshold = \alpha + k \times \sigma \times \left(1 - \frac{1}{1 + \sigma^2}\right)$$

where $k \in \{3\}$ defines high cut-offs. α and σ are considered to be the mean and standard deviation of ECC/edge weight values. Once the target neighborhood network is refined by double filtering, functions of target proteins are predicted using protein sequence, protein domain, and PPI network separately, which will be discussed in the upcoming sections.

Sequence-based prediction

In this section, the functions of target proteins are predicted using protein sequences. Since proteins are formed of amino acids, protein sequence always plays a significant role in target protein function prediction. Sequence-based information is extracted by computing and assigning a Physico-chemical property score to all proteins in the target neighborhood graph, including the target itself. Physico-chemical property score (Jiang and McQuay, 2011) is the average of the values obtained from various Physico-chemical properties of protein/amino acid sequence. In this proposed work, seven Physico-chemical properties are considered, which are:

- *Extinction Coefficient (Eprotein)*
- *Absorbance (Optical Density)*
- *Number of Negatively Charged Residues (Nneg)*
- *Number of Positively Charged Residues (Npos)*
- *Aliphatic Index (AI)*
- *IP/mol weight*
- *Hydrophobicity (Hphb)*

Initially, node degree is computed for each member belonging to the refined neighborhood network for each target protein. The node degrees are then sorted in descending order. Now protein clusters are formed for each target protein. The protein with the highest node degree is selected as the first seed of the cluster. Then the distance between the seed and other proteins in the neighborhood of each target protein is computed based on the Euclidean distance. The Physico-chemical property score serves as an input to the Euclidean distance. If the distance is less than a specific threshold, then the inter-connected protein of the seed is incorporated in the cluster and is removed from the node

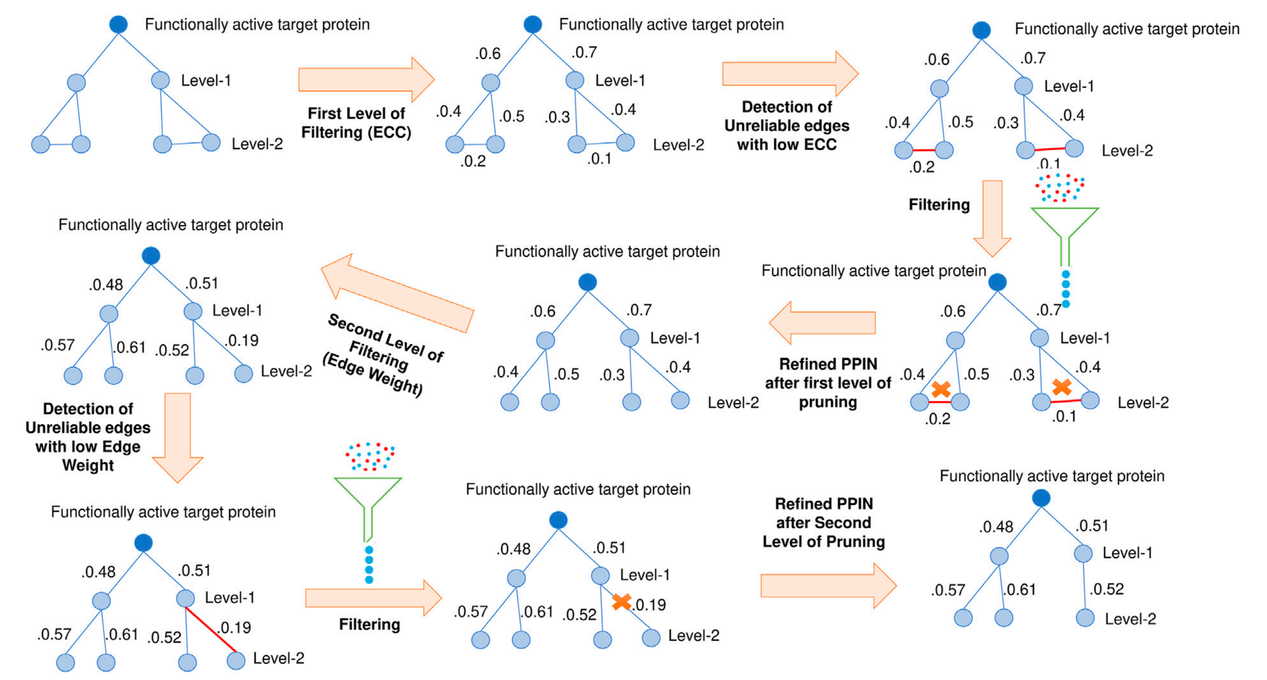

Functionally active target protein
Level-1
Level-2
First Level of Filtering (ECC)
Functionally active target protein
.6
.7
Level-1
.4
.5
.3
.4
.2
.1
Level-2
Detection of Unreliable edges with low ECC
Functionally active target protein
.6
.7
Level-1
.4
.5
.3
.4
.2
.1
Level-2
Filtering
Functionally active target protein
.6
.7
Level-1
.4
.5
.3
.4
.2
.1
Level-2
Refined PPIN after first level of pruning
Functionally active target protein
.6
.7
Level-1
.4
.5
.3
.4
Level-2
Second Level of Filtering (Edge Weight)
Functionally active target protein
.48
.51
Level-1
.57
.61
.52
.19
Level-2
Detection of Unreliable edges with low Edge Weight
Functionally active target protein
.48
.51
Level-1
.57
.61
.52
.19
Level-2
Filtering
Functionally active target protein
.48
.51
Level-1
.57
.61
.52
.19
Level-2
Refined PPIN after Second Level of Pruning
Functionally active target protein
.48
.51
Level-1
.57
.61
.52
Level-2

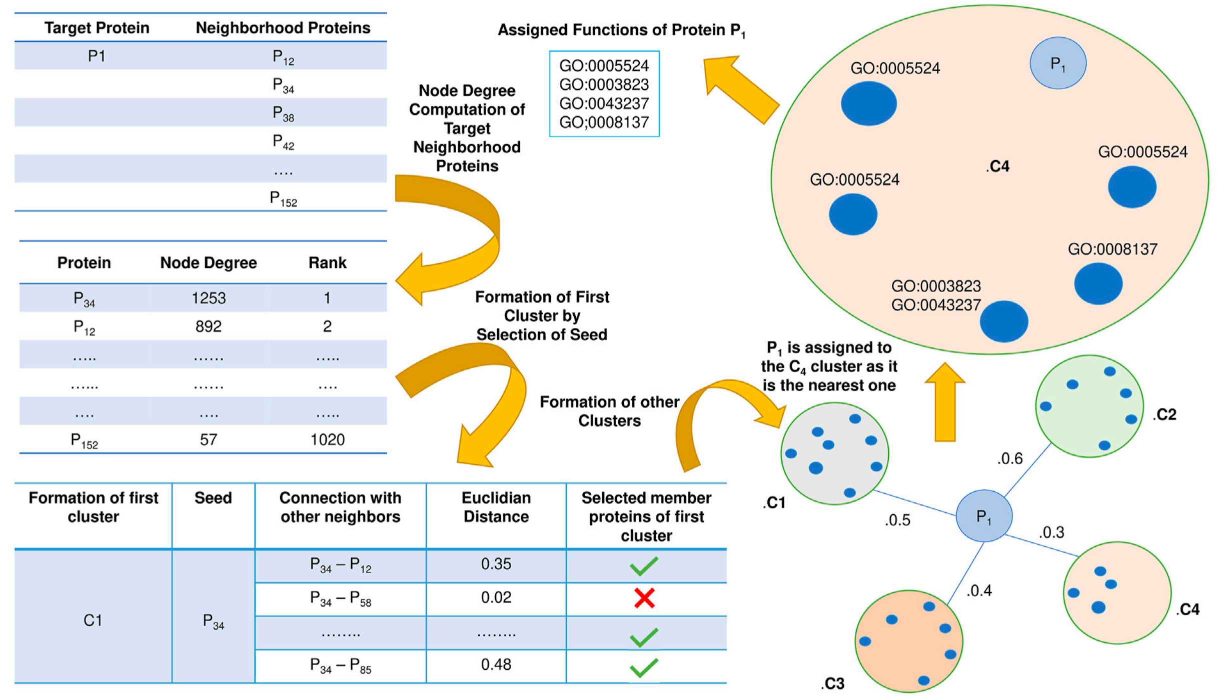
FIGURE 3
Double filtering of target neighborhood graph. Edge clustering coefficient and edge weight are computed based on which non-essential nodes are filtered.

degree list. Then the next node with the highest degree is considered a seed, and its corresponding clusters are formed similarly to the previous one. Thus, the clusters created are validated using inter-cluster and intra-cluster distances so that no miss classification or overlap is present.

Once the cluster formation in each target protein's neighborhood is finished, each cluster's Physico-chemical property score is evaluated. Physico-chemical property score of a cluster is nothing but the average of the earlier computed Physico-chemical score of each constituent protein in the corresponding cluster. Then the Euclidean distance of the Physico-chemical property score between each target protein and its corresponding neighborhood clusters is calculated, and the target protein is assigned to the nearest cluster. Now the cluster contains more than one protein. So, it is not desirable that the functions of all the existing proteins in the chosen cluster are assigned to the target since it will enhance the false positives leading to an abrupt fall in the prediction accuracy level. Considering this fact, intra-cluster distance based on the Euclidean distance of the Physico-chemical property score is computed between the target and the other remaining proteins in the corresponding cluster. Functions of the selected protein having the least distance are assigned to the target protein. The schematic diagram of the entire sequence-based prediction is shown in Figure 4.

Domain-domain interaction-based prediction

Protein domains are the independent units that are responsible for protein function. The study of domain-domain interaction may lead to better protein function prediction. So, this proposed methodology uses PFP-GO protein domains for target protein function prediction. For each node in the refined neighborhood graph of the target protein, its STRING-id is fetched from the STRING database. Each of these STRING-ids is mapped to the Uniport database to obtain its corresponding Uniprot-id. These Uniprot-ids are mapped to their respective PFAM entries (ids), which are used to fetch the PFAM domains using the PFAM database. These derived PFAM domains of the neighborhood graph of each target protein are also checked and validated using the DOMINE database. All the mappings are considered if the one-to-many mapping occurs during the linking as mentioned earlier between databases. Once all the PFAM domains are obtained, the GO terms (protein functions) corresponding to these interacting domains are deduced from the GO Consortium database. Each of these GO terms is assigned a particular ranking based on the frequency of their occurrence. GO terms with the highest ranking are back propagated from the bottom to the top and allocated to the target protein. The entire schematic diagram for this phase has been highlighted in Figure 5.



Target Protein	Neighborhood Proteins
P1	P_{12}
	P_{34}
	P_{38}
	P_{42}
	….
	P_{152}

Protein	Node Degree	Rank
P_{34}	1253	1
P_{12}	892	2
…..	……	…..
……	……	….
….	….	…..
P_{152}	57	1020

Formation of first cluster	Seed	Connection with other neighbors	Euclidian Distance	Selected member proteins of first cluster
C1	P_{34}	$P_{34} – P_{12}$	0.35	✓
		$P_{34} – P_{58}$	0.02	✗
		……..	……..	✓
		$P_{34} – P_{85}$	0.48	✓

Node Degree Computation of Target Neighborhood Proteins
Formation of First Cluster by Selection of Seed
Formation of other Clusters
Assigned Functions of Protein P_1
GO:0005524
GO:0003823
GO:0043237
GO;0008137
P_1
GO:0005524
GO:0005524
.C4
GO:0005524
GO:0008137
GO:0003823
GO:0043237
P_1 is assigned to the C_4 cluster as it is the nearest one
.C1
.C2
.C3
.C4
P_1
.0.6
.0.5
.0.3
.0.4

FIGURE 4
Schematic representation of Sequence-based protein function prediction. The essential aspects of this phase are the selection of seeds followed by the formation of clusters and computation of Physico-chemical properties.

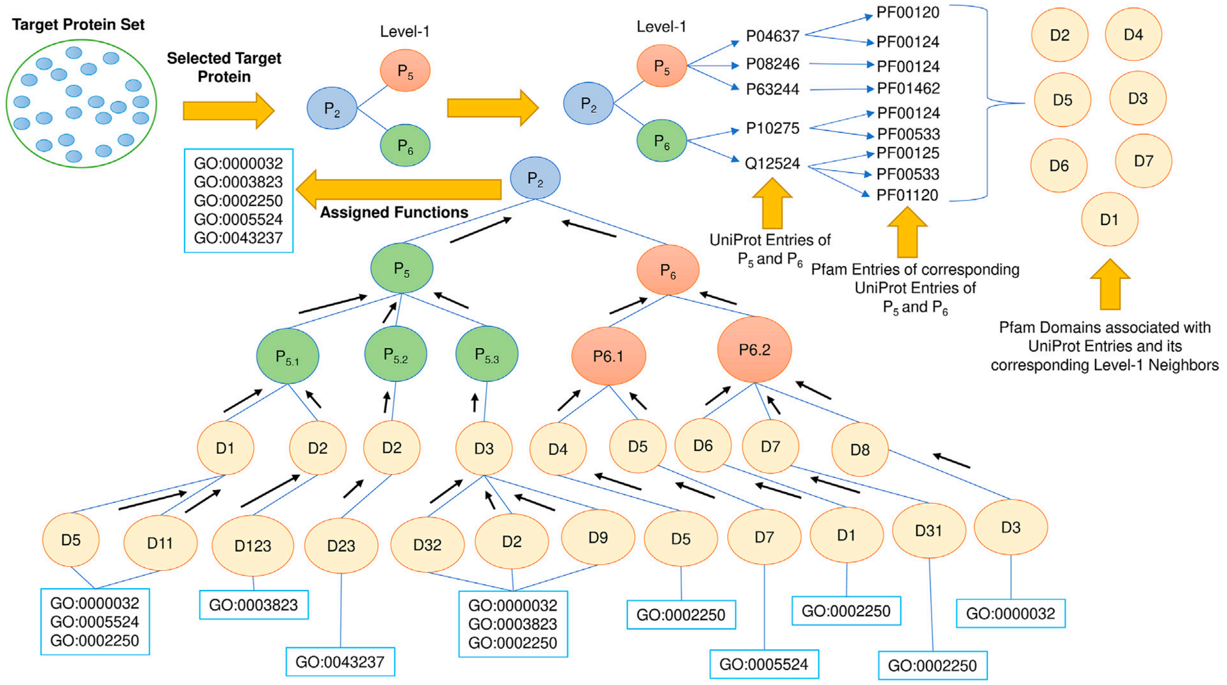
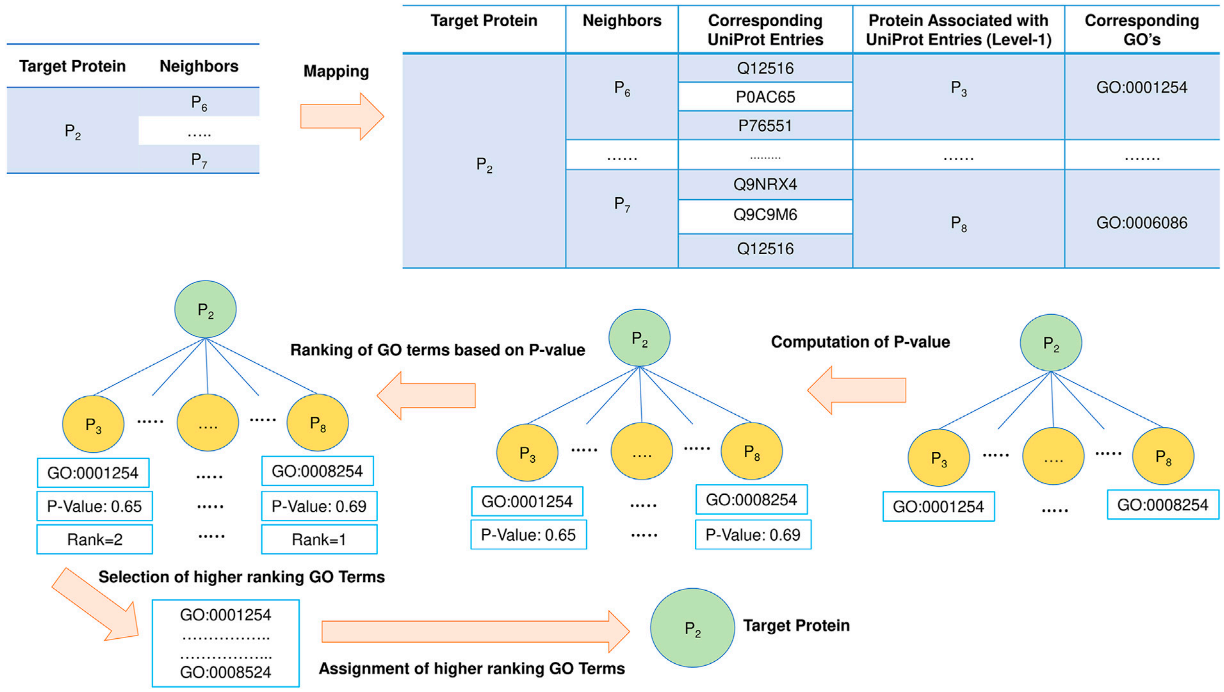


FIGURE 5
Working strategy of Domain-based protein function prediction. Four databases, String, UniProt, PFAM, and DOMINE, are used in this phase.



Target Protein	Neighbors
P_2	P_6
	.....
	P_7

Target Protein	Neighbors	Corresponding UniProt Entries	Protein Associated with UniProt Entries (Level-1)	Corresponding GO's
P_2	P_6	Q12516	P_3	GO:0001254
		P0AC65		
		P76551		
	......	....	......	....
	P_7	Q9NRX4	P_8	GO:0006086
		Q9C9M6		
		Q12516		

FIGURE 6
PPI network-based protein function prediction. GO term enrichment with p value is vital in this phase.

Topology or neighborhood-based prediction from protein-protein interaction network

The functionality of a protein is governed by its topological position in the network. The proteins in the densely related subgraph in the PPI network tend to be more significant in disease propagation and drug targets. This is the most important aspect for which the neighborhood graph of each target protein in PFP-GO gets pruned and filtered before proceeding with the function prediction of the target proteins. In this section, at first, Uniprot-id from Uniprot are fetched from the corresponding STRING-id of each level-1 protein in the pruned neighborhood graph of the target protein. The associated GO terms and proteins (considered level-1 of the target) with these Uniprot-ids are derived from the Uniprot database. The same is also implemented for level-2 of the target protein (Sengupta et al., 2018). Once all the GO terms are fetched, each of them is ranked by computing its enrichment in the PPI network using the p value with Fisher's exact test (Piovesan et al., 2015). Top-ranked GO terms are finally allocated to the target protein. The schematic diagram for this phase of PPI network-topology-based prediction is shown in Figure 6.

Integrated prediction using sequence, domain, and protein-protein interaction

The consensus technique is more effective if the results from orthogonal sources are merged to generate consistent results. In PFP-GO, PPI network, domain, and sequence data are used. For the final prediction, the n-star consensus method is used. As discussed earlier, the proposed method uses a 3-star consensus between three different predictors. 1-star consensus assigns those GO terms predicted by at least one of the predictors to the target protein. The 1-star is the least reliable as it gives the maximum number of GO terms. The 2-star consensus assigns the GO terms commonly predicted by at least two different predictors to the target protein. In contrast, the 3-star consensus (Chatterjee et al., 2016) is the most reliable as it only assigns GO terms commonly predicted by all three predictors for the target protein.

Results

PFP-GO first selects a functionally active protein set from the database, considered target proteins. It then predicts the functions of the target proteins using sequence-based, domain-based, and neighborhood Protein Interaction based approaches discussed earlier in the methodology section. The

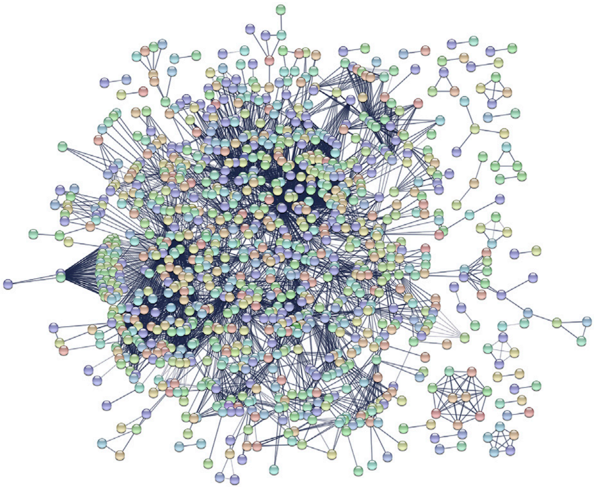


FIGURE 7
Sample human PPI network from STRING database. It consists of nodes and interactions between them.

proposed method fetches the proteins in the String database for the target set selection. These proteins are mapped to the UniProt database. The STRING database contains 19,247 human proteins mapped to the UniProt database to retrieve the associated GO terms. PFP-GO utilizes the mapping of UniProt to retrieve GO terms because String database entries are not directly associated with GO annotation. In this proposed work, the human PPI network of String databases is used because it has 85, 58,002 interactions which is significantly higher than UniProt, which has only 46,410 interactions. A network diagram of the Human String database consisting of 2000 nodes is highlighted in Figure 7. Since the mapping between String and UniProt is one-to-many, each string entry has one or more UniProt IDs. However, suppose the mapping is not present for a particular protein. In that case, the corresponding sequence of String data is fetched, and the proteins having 90% similarity with it are considered from the UniProt database. Their corresponding GO terms are also used for further experimentation. In the UniProt database, out of 12,366 GO terms, 1547 GO terms fall under the Cellular Components category, while 4,105 and 11,263 GO terms are classified under Molecular Function and Biological Process, respectively. Then the frequency of each GO term is calculated, and the top 10 GO terms are fetched along with their UniProt IDs. PFP-GO detects 9,141 proteins associated with the top 10 GO terms, which is reverse mapped to the STRING dataset to get 6,999 proteins. Then it selects a random 20% of proteins out of 6,999, which is near about 1,400 proteins. From these 1,400 proteins, 639 unique proteins are finally filtered out after redundancy removal. These proteins are ultimately considered functionally active targets.

TABLE 2 Performance Analysis of INGA and PFP-GO based on PPI network, sequence, and domain.

Methodology	Precision	Recall	F-score
PFP-GO	0.67	0.58	0.62
INGA Piovesan et al., (2015)	0.44	0.51	0.47

The performance measure of PFP-GO is evaluated using Precision (P), Recall (R), and F-Score (F). PFP-GO is a combined methodology based on three heterogeneous resources, i.e., protein sequence, protein domain, and PPI network. It achieves an overall precision, recall, and F-score of 0.67, 0.58, and 0.62, respectively. It is initially compared with INGA (Piovesan et al., 2015), as shown in Table 2, since it uses the same heterogeneous resources as that PFP-GO. However, INGA lacks proper filtering and pruning of the neighborhood graph of the target protein. Besides consideration of Physico-chemical properties of protein sequence in PFP-GO instead of just applying sequence comparison algorithm to estimate sequence similarity is another aspect for outperforming INGA.

The same has also been compared with four of the existing methods: FunApriori (Prasad et al., 2017), the neighborhood counting method (Schwikowski et al., 2000), the chi-square method (Hishigaki et al., 2001), a recent version of the neighbor relativity coefficient (NRC) (Moosavi et al., 2013), FS-weight based method (Chua et al., 2006). It should be noted here that all these methods are based on the PPI network alone. To remove biases and to compare in a common field, performance analysis of PFP-GO is estimated on the prediction of the PPI network only. The performance is highlighted in Table 3.

The major limitation of the chi-square method is that it is suitable only for the denser part of the network. Thus, network sparseness may lead to the degradation of performance evaluation compared to the others. The inclusion of level-1 and level-2 neighbors increases the accuracy rates in all except chi-square #1 and FS-weight #1 (using only the first level).

The neighborhood counting method is simple but still lags more like NRC and FS-weight #1 and #2 (using both levels) as it fails to differentiate between them. Although NRC and FunApriori perform better than the other, they fall behind PFP-GO since it does not focus on eliminating non-essential proteins from the target neighborhood.

PFP-GO based on only PPI network and sequence is also tested against similar kinds of existing methodologies like NAIVE (Murphy, 2006) and BLAST (Mount, 2007) method as reported in (Piovesan et al., 2015), Multi-Label Protein Function Prediction (ML_PFP) (Saha, Prasad, et al., 2018) and DeepGO (Kulmanov et al., 2018). Table 4 highlights the

TABLE 3 Performance analysis of PFP-GO with other methods based on PPI network.

Methodology	Precision	Recall	F-score
PFP-GO	0.74	0.67	0.73
FunApriori Prasad et al., (2017)	0.57	0.61	0.58
Chi square #1and2 Hishigaki et al., (2001)	0.13	0.12	0.12
Chi square #1 Hishigaki et al., (2001)	0.12	0.15	0.13
Neighborhood counting #1and2 Schwikowski et al., (2000)	0.21	0.25	0.18
Neighborhood counting #1 Schwikowski et al., (2000)	0.15	0.21	0.17
Fs-weight #1and2 Chua et al., (2006)	0.24	0.22	0.22
Fs-weight #1 Chua et al., (2006)	0.16	0.19	0.19
Nrc Moosavi et al., (2013)	0.25	0.24	0.22

TABLE 4 Performance analysis of PFP-GO with other methods based on PPI network and sequence.

Methodology	Precision	Recall	F-score
PFP-GO	0.52	0.64	0.56
Deep_GO Kulmanov et al., (2018)	0.48	0.49	0.48
BLAST Mount, (2007); Piovesan et al., (2015)	0.30	0.50	0.37
NAÏVE Murphy, (2006); Piovesan et al., (2015)	0.33	0.31	0.31

TABLE 5 Performance analysis of INGA and PFP-GO separately on CC, MF and BP.

	Precision			Recall			F-score		
Methodology	BP	MF	CC	BP	MF	CC	BP	MF	CC
PFP-GO	0.49	0.51	0.48	0.95	0.98	0.95	0.64	0.67	0.64
INGA Piovesan et al., (2015)	0.37	0.53	0.42	0.33	0.63	0.63	0.58	0.57	0.49

TABLE 6 Performance analysis of PFP-GO with other methods based on Fmax score.

Methodology	BP	MF	CC
PFP-GO	0.65	0.61	0.66
NetGO 3.0 You et al., (2019)	0.64	0.43	0.66
Deep_GO_Plus Kulmanov and Hoehndorf, (2020)	0.57	0.41	0.59
BLAST Mount, (2007); Piovesan et al., (2015)	0.63	0.31	0.56
NAÏVE Murphy, (2006); Piovesan et al., (2015)	0.4	0.23	0.54

TABLE 7 Top-ranked gene ontology terms selected from GGA validation.

Rank	Gene	GO-terms
1	ENSG00000123131	GO:0000049
2	ENSG00000123131	GO:0001731
3	ENSG00000123131	GO:0003743
4	ENSG00000130741	GO:0005576
5	ENSG00000130741	GO:0005634
6	ENSG00000130741	GO:0005783

entire scenario. Moreover, the prediction performance of PFP-GO has also been evaluated with INGA separately on GO terms: Cellular Components (CC), Molecular Functions (MF) and Biological Process (BP), highlighted in Table 5. F_{MAX} score for BP, MF and CC has also been taken into account, and the same is compared with other methods like NetGO 3.0 (You et al., 2019), DeepGOPlus (Kulmanov and Hoehndorf, 2020), BLAST (Mount, 2007) and NAÏVE (Murphy, 2006). The result is displayed in Table 6. None of these methods considers filtering or including various sequence-derived features like an aliphatic index, etc., because they fail to perform better than PFP-GO. Moreover, Moreover, DeepGO cannot predict protein functions with a sequence length >1,000, which is another snag. ML_PFP has used protein sequence and PPI network quite effectively but uses only edge weight as the only parameter for screening the non-reliable edges in PPIN. In contrast, PFP-GO uses 2-pass filtering and pruning approach.

To further validate the predicted and ranked Gene Ontology terms, we used the meta-network created in (Halder et al., 2020) and (Chiliński et al., 2021) to study the importance of some of the GO terms in the perspective of the 3D-structure of the genome. To include this assessment, we create a Gene-Gene association network following the ideas presented in the work of Chiliński et al. (2021); Halder et al. (2020). We derive the Genomic association network from the 3D chromatin structure. After we created the networks, we mapped the unknown test set to the network and found the level-1 and level-2 neighboring genes from each target node. We end up with a similar tree, as shown in Figure 5. However, here the nodes in the tree are genes. Then we map the genes to the GO terms from the leaf nodes and propagate them to the target node (Sengupta et al., 2018). Once all the GO terms are fetched, each of them is ranked by computing its enrichment in the PPI network using the p value with Fisher's exact test (Piovesan et al., 2015). Top-ranked GO terms are finally allocated to the target protein. From the ranking, we obtained the top Gene ontology terms, which are displayed in Table 7.

Conclusion

From Table 2, Table 3, and Table 4, it can be inferred that our method PFP-GO outperforms the other methodologies in the same dataset of humans in terms of precision, recall, and F-score values due to several reasons: 1) The target set of proteins are selected from high-ranking GO terms which implies the fact that only proteins having high connectivity are involved. 2) Pruning and double filtering of proteins are executed by eliminating Bridge, Shore, and Fjord proteins (non-essential proteins) which have not been taken into account by the other methods, which is the primary cause for the increase in false rates. 3) Besides consideration of every GO and non-GO term, prediction provides a proper equilibrium in the proposed methodology. 4) Moreover, PFP-GO combines prediction from three orthogonal sources, i.e., sequence-based predictor, domain interaction network-based predictor, and protein interaction network-based predictor, to predict the protein's function. All these lead to the enhancement of our prediction accuracy.

Besides, it should be noted that PFP-GO can perform better whether it considers the PPI network alone, PPI network and sequence, or the combination of the trio: PPI network, sequence, and domain. It can also identify functionally active proteins which may be transmitted in identifying possible drug targets in the future (Anup Kumar Halder et al., 2018; Saha, Sengupta, et al., 2018). Recently our work has been limited to human-specific datasets, which can be extended further to other organisms. The PFP-GO software package and the complete source code are available in the public domain for noncommercial research use at https://sites.google.com/view/pfp-go/.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Author contributions

KS and SS proposed the idea of protein function prediction by integrating various sources of information. KS, SS, PC, AH, and SB developed further the initial concept and performed the whole study. KS, SS, PC, and AH gathered the data, implemented the algorithms and performed the simulations, and performed the statistical analysis. KS, SS, PC, AH, MN, SB, and DP prepared the manuscript. MN, SB, and DP consulted and corrected the research and the manuscript. All authors approved the final manuscript. KS and SS contributed equally to the article.

Funding

This project is partially supported by the CMATER research laboratory of the Computer Science and Engineering Department, Jadavpur University, India, and UGC Research Award [F.30-31/2016(SA-II)] from UGC, Government of India, and DBT project (No. BT/PR16356/BID/7/596/2016), Ministry of Science and Technology, Government of India. KS is thankful to IDUB grant BOB-IDUB-622-197/2021. This work has been supported by the Polish National Science Centre (2019/35/O/ST6/02484 and 2020/37/B/NZ2/03757), Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund (TEAM to DP). The work has been co-supported by European Commission Horizon 2020 Marie Skłodowska-Curie ITN Enpathy grant "Molecular Basis of Human enhanceropathies;" and National Institutes of Health United States 4DNucleome grant 1U54DK107967-01 "Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation." DP's effort has been co-supported as RENOIR Project by the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 691152 and by Ministry of Science and Higher Education (Poland), Grant Nos. W34/H2020/2016, 329025/PnH

/2016. DP's efforts have been co-funded by Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme. Computations have been performed thanks to the Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology using Artificial Intelligence HPC platform financed by Polish Ministry of Science andHigher Education (decision no. 7054/IA/SP/2020 of 2020-08-28).

Acknowledgments

Authors are thankful to the CMATER research laboratory of the Computer Science Department, Jadavpur University, India, for providing infrastructure facilities during progress of the work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215 (3), 403–410. doi:10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25 (17), 3389–3402. doi:10.1093/nar/25.17.3389

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* 181 (4096), 223–230. doi:10.1126/science.181.4096.223

Ashburner, M., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: Tool for the unification of biology. The gene ontology Consortium. *Nat. Genet.* 25 (1), 25–29. doi:10.1038/75556

Attwood, T. K. (2002). The PRINTS database: A resource for identification of protein families. *Brief. Bioinform.* 3 (3), 252–263. doi:10.1093/bib/3.3.252

Bader, G. D., and Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinforma.* 4 (1), 2. doi:10.1186/1471-2105-4-2

Bao, W., Cui, Q., Chen, B., and Yang, B. (2022). Phage_UniR_LGBM: Phage virion proteins classification with UniRep features and LightGBM model. *Comput. Math. Methods Med.* 2022, 9470683. doi:10.1155/2022/9470683

Bao, W., Yang, B., and Chen, B. (2021). 2-hydr_ensemble: Lysine 2-hydroxyisobutyrylation identification with ensemble method. *Chemom. Intelligent Laboratory Syst.* 215, 104351. doi:10.1016/j.chemolab.2021.104351

Bao, W., Yuan, C.-A., Zhang, Y., Han, K., Nandi, A. K., Honig, B., et al. (2017). Mutli-features prediction of protein translational modification sites. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15 (5), 1453–1460. doi:10.1109/TCBB.2017.2752703

Chatterjee, P., Basu, S., Kundu, M., Nasipuri, M., and Plewczynski, D. (2011). PPI_SVM: Prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables. *Cell. Mol. Biol. Lett.* 16 (2), 264–278. doi:10.2478/s11658-011-0008-x

Chatterjee, P., Basu, S., Zubek, J., Kundu, M., Nasipuri, M., and Plewczynski, D. (2016). PDP-CON: Prediction of domain/linker residues in protein sequences using a consensus approach. *J. Mol. Model.* 22 (4), 72–15. doi:10.1007/s00894-016-2933-0

Chen, J., Hsu, W., Lee, M. L., and Ng, S.-K. (2007). Labeling network motifs in protein interactomes for protein function prediction. *IEEE 23rd Int. Conf. Data Eng.* 2007, 546–555. doi:10.1109/ICDE.2007.367900

Chen, Y., and Xu, D. (2004). Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 32 (21), 6414–6424. doi:10.1093/nar/gkh978

Chiliński, M., Sengupta, K., and Plewczynski, D. (2021). From DNA human sequence to the chromatin higher order organisation and its biological meaning: Using biomolecular interaction networks to understand the influence of structural variation on spatial genome organisation and its functional effect. *Seminars Cell & Dev. Biol.* 121, 171–185. doi:10.1016/j.semcdb.2021.08.007

Chitale, M., Hawkins, T., and Kihara, D. (2009). Automated prediction of protein function from sequence. *Predict. Protein Strucutre, Funct. Interact.*, 63–86. doi:10.1002/9780470741894.ch3

Chua, H. N., Sung, W.-K., and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics* 22 (13), 1623–1630. doi:10.1093/bioinformatics/btl145

Consortium, G. O. (2018). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47 (1), D330–D338. doi:10.1093/nar/gky1055

Consortium, U. (2015). UniProt: A hub for protein information. *Nucleic Acids Res.* 43 (1), D204–D212. doi:10.1093/nar/gku989

Corpet, F., Servant, F., Gouzy, J., and Kahn, D. (2000). ProDom and ProDom-CG: Tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* 28 (1), 267–269. doi:10.1093/nar/28.1.267

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: A fingerprint of proteins that physically interact. *Trends biochem. Sci.* 23 (9), 324–328. doi:10.1016/s0968-0004(98)01274-2

Deng, M., Mehta, S., Sun, F., and Chen, T. (2002). Inferring domain – domain interactions from protein – protein interactions. *Genome Res.* 12, 1540–1548. doi:10.1101/gr.153002.2

Deng, M., Sun, F., and Chen, T. (2002). "Assessment of the reliability of protein-protein interactions and protein function prediction," in *Biocomputing 2003* (World Scientific), 140–151.

Deng, M., Tu, Z., Sun, F., and Chen, T. (2004). Mapping gene ontology to proteins based on protein–protein interaction data. *Bioinformatics* 20 (6), 895–902. doi:10.1093/bioinformatics/btg500

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* 44 (D1), D279–D285. doi:10.1093/nar/gkv1344

Forslund, K., and Sonnhammer, E. L. L. (2008). Predicting protein function from domain content. *Bioinformatics* 24 (15), 1681–1687. doi:10.1093/bioinformatics/btn312

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., et al. (2013). STRING v9. 1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–D815. doi:10.1093/nar/gks1094

Garg, A., Bhasin, M., and Raghava, G. P. S. (2005). Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J. Biol. Chem.* 280 (15), 14427–14432. doi:10.1074/jbc.M411789200

Gherardini, P. F., and Helmer-Citterich, M. (2008). Structure-based function prediction: Approaches and applications. *Brief. Funct. Genomic. Proteomic.* 7 (4), 291–302. doi:10.1093/bfgp/eln030

Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., et al. (2021). Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* 12 (1), 3168. doi:10.1038/s41467-021-23303-9

Halder, A. K., Dutta, P., Kundu, M., Basu, S., and Nasipuri, M. (2018). Review of computational methods for virus – host protein interaction prediction : A case study on novel ebola – human interactions. *Brief. Funct. Genomics* 17 (2017), 381–391. doi:10.1093/bfgp/elx026

Halder, A. K., Chatterjee, P., Nasipuri, M., Plewczynski, D., and Basu, S. (2019). 3gClust: Human protein cluster Analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16 (6), 1773–1784. doi:10.1109/TCBB.2018.2840996

Halder, A. K., Denkiewicz, M., Sengupta, K., Basu, S., and Plewczynski, D. (2020). Aggregated network centrality shows non-random structure of genomic and proteomic networks. *Methods* 181, 5–14. doi:10.1016/j.ymeth.2019.11.006

Hanna, E. M., and Zaki, N. (2014). Detecting protein complexes in protein interaction networks using a ranking algorithm with a refined merging procedure. *BMC Bioinforma.* 15 (1), 204–211. doi:10.1186/1471-2105-15-204

Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., and Takagi, T. (2001). Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast* 18 (6), 523–531. doi:10.1002/yea.706

Huang, Y., and Li, Y. (2004). Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics* 20 (1), 21–28. doi:10.1093/bioinformatics/btg366

Jiang, J. Q., and McQuay, L. J. (2011). Predicting protein function by multi-label correlated semi-supervised learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9 (4), 1059–1069. doi:10.1109/TCBB.2011.156

Karaoz, U., Murali, T. M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C. R., et al. (2004). Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl. Acad. Sci. U. S. A.* 101 (9), 2888–2893. doi:10.1073/pnas.0307326101

Kihara, D. (2011). *Protein function prediction for omics era*. Germany: Springer.

King, A. D., Pržulj, N., and Jurisica, I. (2004). Protein complex prediction via cost-based clustering. *Bioinformatics* 20 (17), 3013–3020. doi:10.1093/bioinformatics/bth351

Kulmanov, M., and Hoehndorf, R. (2020). DeepGOPlus: Improved protein function prediction from sequence. *Bioinformatics* 36 (2), 422–429. doi:10.1093/bioinformatics/btz595

Kulmanov, M., Khan, M. A., Hoehndorf, R., and Wren, J. (2018). DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* 34 (4), 660–668. doi:10.1093/bioinformatics/btx624

Letovsky, S., and Kasif, S. (2003). Predicting protein function from protein/protein interaction data: A probabilistic approach. *Bioinformatics* 19 (1), i197–i204. doi:10.1093/bioinformatics/btg1026

Li, M., Shi, W., Zhang, F., Zeng, M., and Li, Y. (2022). A deep learning framework for predicting protein functions with co-occurrence of GO terms. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 27, 1. doi:10.1109/TCBB.2022.3170719

Li, Z., Liao, B., Li, Y., Liu, W., Chen, M., and Cai, L. (2018). Gene function prediction based on combining gene ontology hierarchy with multi-instance multi-label learning. *RSC Adv.* 8 (50), 28503–28509. doi:10.1039/c8ra05122d

Moosavi, S., Rahgozar, M., and Rahimi, A. (2013). Protein function prediction using neighbor relativity in protein–protein interaction network. *Comput. Biol. Chem.* 43, 11–16. doi:10.1016/j.compbiolchem.2012.12.003

Mount, D. W. (2007). Using the basic local alignment search tool (BLAST). *Cold Spring Harb. Protoc.* 2007 (7), pdb.top17–top17. doi:10.1101/pdb.top17

Murphy, K. P. (2006). Naive Bayes classifiers. *Univ. B. C.* 18 (60), 1–8.

Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 21 (1), i302–i310. doi:10.1093/bioinformatics/bti1054

Najmanovich, R. J., Torrance, J. W., and Thornton, J. M. (2005). Prediction of protein function from structure: Insights from methods for the detection of local structural similarities. *Biotechniques* 38 (6), 847, 849, 851–851. doi:10.2144/05386TE01

Nielsen, H., Engelbrecht, J., Brunak, S., and Von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10 (1), 1–6. doi:10.1093/protein/10.1.1

Pandey, G., Kumar, V., and Steinbach, M. (2006). Computational approaches for protein function prediction: A survey. *Digital Conservancy. https://hdl.handle.net/11299/215713.*

Pandit, S. B., Gosar, D., Abhiman, S., Sujatha, S., Dixit, S. S., Mhatre, N. S., et al. (2002). SUPFAM—a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: Implications for structural genomics and function annotation in genomes. *Nucleic Acids Res.* 30 (1), 289–293. doi:10.1093/nar/30.1.289

Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Sci.* 4 (6), 1145–1160. doi:10.1002/pro.5560040613

Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* 85 (8), 2444–2448. doi:10.1073/pnas.85.8.2444

Pearson, W. R., and Sierk, M. L. (2005). The limits of protein sequence comparison? *Curr. Opin. Struct. Biol.* 15 (3), 254–260. doi:10.1016/j.sbi.2005.05.005

Peng, W., Wang, J., Cai, J., Chen, L., Li, M., and Wu, F.-X. (2014). Improving protein function prediction using domain and protein complexes in PPI networks. *BMC Syst. Biol.* 8 (1), 35. doi:10.1186/1752-0509-8-35

Peng, W., Wang, J., Wang, W., Liu, Q., Wu, F.-X., and Pan, Y. (2012). Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. *BMC Syst. Biol.* 6 (1), 87–17. doi:10.1186/1752-0509-6-87

Pietrokovski, S., Henikoff, J. G., and Henikoff, S. (1996). The blocks database—A system for protein classification. *Nucleic Acids Res.* 24 (1), 197–200. doi:10.1093/nar/24.1.197

Piovesan, D., Giollo, M., Leonardi, E., Ferrari, C., and Tosatto, S. C. E. (2015). Inga: Protein function prediction combining interaction networks, domain assignments and sequence similarity. *Nucleic Acids Res.* 43 (1), W134–W140. doi:10.1093/nar/gkv523

Prasad, A., Saha, S., Chatterjee, P., Basu, S., and Nasipuri, M. (2017). Protein function prediction from protein interaction network using bottom-up L2L apriori algorithm. *Int. Conf. Comput. Intell. Commun. Bus. Anal.*, 3–16. doi:10.1007/978-981-10-6430-2_1

Rentzsch, R., and Orengo, C. A. (2013). Protein function prediction using domain families. *BMC Bioinforma.* 14 (3), S5–S14. doi:10.1186/1471-2105-14-S3-S5

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., et al. (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.* 32 (18), 5539–5545. doi:10.1093/nar/gkh894

Saha, S., Chatterjee, P., Basu, S., Kundu, M., and Nasipuri, M. (2014). FunPred-1: Protein function prediction from a protein interaction network using neighborhood analysis. *Cell. Mol. Biol. Lett.* 19 (4), 675–691. doi:10.2478/s11658-014-0221-5

Saha, S., Prasad, A., Chatterjee, P., Basu, S., and Nasipuri, M. (2018). Protein function prediction from protein–protein interaction network using gene ontology based neighborhood analysis and physico-chemical features. *J. Bioinform. Comput. Biol.* 16 (06), 1850025. doi:10.1142/s0219720018500257

Saha, S., Sengupta, K., Chatterjee, P., Basu, S., and Nasipuri, M. (2018). Analysis of protein targets in pathogen-host interaction in infectious diseases: A case study on plasmodium falciparum and *Homo sapiens* interaction network. *Brief. Funct. Genomics* 17 (6), 441–450. doi:10.1093/bfgp/elx024

Sarda, D., Chua, G. H., Li, K.-B., and Krishnan, A. (2005). pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. *BMC Bioinforma* 6 (1), 152–212. doi:10.1186/1471-2105-6-152

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein–protein interactions in yeast. *Nat. Biotechnol.* 18 (12), 1257–1261. doi:10.1038/82360

Sengupta, K., Saha, S., Chatterjee, P., Kundu, M., Nasipuri, M., and Basu, S. (2018). "Ranked gene ontology based protein function prediction by analysis of protein–protein interactions," in *Information and decision sciences* (Germay: Springer), 419–427.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Mol. Syst. Biol.* 3 (1), 88. doi:10.1038/msb4100129

Spirin, V., and Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U. S. A.* 100 (21), 12123–12128. doi:10.1073/pnas.2032324100

Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.* 21 (6), 697–700. doi:10.1038/nbt825

Wang, P., and Xiao, X. (2014). NRPred-FS: A feature selection based two level predictor for nuclear receptors. *J. Proteomics Bioinform.*, s9. *supplement 9article 002.* doi:10.4172/jpb.s9-002

Wang, S., and Wu, F. (2013). Detecting overlapping protein complexes in PPI networks based on robustness. *Proteome Sci.* 11 (1), S18–8. doi:10.1186/1477-5956-11-S1-S18

Wang, T., and Yang, J. (2010). Predicting subcellular localization of gram-negative bacterial proteins by linear dimensionality reduction method. *Protein Pept. Lett.* 17 (1), 32–37. doi:10.2174/092986610789909494

Wang, Y.-C., Wang, X.-B., Yang, Z.-X., and Deng, N.-Y. (2010). Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. *Protein Pept. Lett.* 17 (11), 1441–1449. doi:10.2174/092986651100901l441

Xiao, X., Wang, P., and Chou, K.-C. (2012). iNR-PhysChem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix. *PloS One* 7 (2), e30869. doi:10.1371/journal.pone.0030869

Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B., and Jothi, R. (2011). Domine: A comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res.* 39 (1), D730–D735. doi:10.1093/nar/gkq1229

You, R., Yao, S., Xiong, Y., Huang, X., Sun, F., Mamitsuka, H., et al. (2019). NetGO: Improving large-scale protein function prediction with massive network information. *Nucleic Acids Res.* 47 (1), W379-W387–W387. doi:10.1093/nar/gkz388

Zhang, F., Song, H., Zeng, M., Li, Y., Kurgan, L., and Li, M. (2019). DeepFunc: A deep learning framework for accurate prediction of protein functions from protein sequences and interactions. *Proteomics* 19 (12), 1900019. doi:10.1002/pmic.201900019

Zhang, F., Song, H., Zeng, M., Wu, F., Li, Y., Pan, Y., et al. (2020). A deep learning framework for gene ontology annotations with sequence-and network-based information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 2208–2217. doi:10.1109/TCBB.2020.2968882

Zhang, X., Wang, L., Liu, H., Zhang, X., Liu, B., Wang, Y., et al. (2021). Prot2GO: Predicting GO annotations from protein sequences and interactions. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 1. doi:10.1109/TCBB.2021.3139841

Zhang, Y., Lin, H., Yang, Z., Wang, J., Liu, Y., and Sang, S. (2016). A method for predicting protein complex in dynamic PPI networks. *BMC Bioinforma.* 17 (7), 229–543. doi:10.1186/s12859-016-1101-y

Zhao, Y., Fu, G., Wang, J., Guo, M., and Yu, G. (2019). Gene function prediction based on gene ontology hierarchy preserving hashing. *Genomics* 111 (3), 334–342. doi:10.1016/j.ygeno.2018.02.008