FunPred 3.0: Improved protein function prediction using protein interaction network

Sovan Saha 1, Piyali Chatterjee 2, Subhadip Basu 3, Mita Nasipuri 3, Dariusz Plewczynski *Corresp. 4, 5

1 Department of Computer Science and Engineering, Dr. Sudhir Chandra Sur Degree Engineering College, Kolkata, West Bengal, India
2 Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India
3 Department of Computer Science and Engineering, Jadavpur University, Kolkata, West Bengal, India
4 Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Warsaw, Poland
5 Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

Corresponding Author: Dariusz Plewczynski
Email address: d.plewczynski@cent.uw.edu.pl

Proteins are the most versatile macromolecules in living systems and perform crucial biological functions. In the advent of the post genomic era, the next generation sequencing is done routinely at the population scale for a variety of species. The challenging problem is to massively determine the functions of proteins that are yet not characterized by detailed experimental studies. Identification of protein functions experimentally is a laborious and time-consuming task involving many resources. We therefore propose the automated protein function prediction methodology using *in silico* algorithms trained on carefully curated experimental datasets. We present the improved protein function prediction tool FunPred 3.0, an extended version of our previous methodology FunPred 2, which exploits neighbourhood properties in protein-protein interaction network (PPIN) and physicochemical properties of amino acids. Our method is validated using the available functional annotations in the PPIN network of *Saccharomyces cerevisiae* in the latest Munich Information Center for Protein (MIPS) dataset. The PPIN data of *Saccharomyces cerevisiae* in MIPS dataset includes 4554 unique proteins in 13528 protein-protein interactions (PPINs) after the elimination of the self-replicating and the self-interacting protein pairs. Using the developed FunPred 3.0 tool, we are able to achieve the mean precision, the recall and the F-score values of 0.55, 0.82 and 0.66 respectively. FunPred 3.0 is then used to predict the functions of unpredicted protein pairs (incomplete and missing functional annotations) in MIPS dataset of *Saccharomyces cerevisiae*. The method is also capable of predicting the subcellular localization of proteins along with its corresponding functions. The code and the complete prediction results are available freely at: https://github.com/SovanSaha/FunPred-3.0.git.
FunPred 3.0: Improved protein function prediction using protein interaction network

Sovan Saha¹, Piyali Chatterjee², Subhadip Basu³, Mita Nasipuri³, Dariusz Plewczynski⁴,⁵

¹ Department of Computer Science and Engineering, Dr. Sudhir Chandra Sur Degree Engineering College, DumDum, Kolkata, India
² Department of Computer Science and Engineering, Netaji Subhash Engineering College, Garia, Kolkata, India
³ Department of Computer Science and Engineering, Jadavpur University, Kolkata, West Bengal, India
⁴ Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Warsaw, Poland
⁵ Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

Corresponding Author:
Dariusz Plewczynski⁴,⁵
Banacha 2c Street, 02-097 Warsaw, Poland
Email address: d.plewczynski@cent.uw.edu.pl
ABSTRACT

Proteins are the most versatile macromolecules in living systems and perform crucial biological functions. In the advent of the post genomic era, the next generation sequencing is done routinely at the population scale for a variety of species. The challenging problem is to massively determine the functions of proteins that are yet not characterized by detailed experimental studies. Identification of protein functions experimentally is a laborious and time-consuming task involving many resources. We therefore propose the automated protein function prediction methodology using in silico algorithms trained on carefully curated experimental datasets. We present the improved protein function prediction tool FunPred 3.0, an extended version of our previous methodology FunPred 2, which exploits neighbourhood properties in protein-protein interaction network (PPIN) and physicochemical properties of amino acids. Our method is validated using the available functional annotations in the PPIN network of Saccharomyces cerevisiae in the latest Munich Information Center for Protein (MIPS) dataset. The PPIN data of Saccharomyces cerevisiae in MIPS dataset includes 4554 unique proteins in 13528 protein-protein interactions (PPINs) after the elimination of the self-replicating and the self-interacting protein pairs. Using the developed FunPred 3.0 tool, we are able to achieve the mean precision, the recall and the F-score values of 0.55, 0.82 and 0.66 respectively. FunPred 3.0 is then used to predict the functions of unpredicted protein pairs (incomplete and missing functional annotations) in MIPS dataset of Saccharomyces cerevisiae. The method is also capable of predicting the subcellular localization of proteins along with its corresponding functions. The code and the complete prediction results are available freely at: https://github.com/SovanSaha/FunPred-3.0.git.

INTRODUCTION

Proteins with similar functions are more likely to interact. If the function of one protein is known then the functions of the binding un-annotated protein may either be experimentally assigned or computationally predicted (Chatterjee et al. 2011a; Chatterjee et al. 2011b; Moosavi et al. 2013; Prasad et al. 2017; Saha et al. 2012; Saha et al. 2014; Sriwastava et al. 2015). Several computational techniques have been developed using either the protein sequence (Ng & Henikoff 2003), protein structure (Lee et al. 2007; Mills et al. 2015), protein-protein interactions (Moosavi et al. 2013; Schwikowski et al. 2000; Vazquez et al. 2003; Xiong et al. 2013), or sequence motifs or signatures (Chatterjee et al. 2011a; Chen et al. 2007; Lichtarge et al. 1996). Protein Interaction datasets are represented as graphs (with every node corresponding to an individual protein and each edge between a pair of nodes representing the interaction between them) can be used to assign biological functions to a protein with an assumption that close neighbors of a protein are functionally similar.
Protein function prediction problem is characterized by several factors like the diversity of members for functional groups, the hierarchical relationships among functional classes, incomplete or missing information about proteins and their functions. Thus, it defines a complex multi-label learning problem (Jiang & McQuay 2012; Valentini 2014; Zhang & Zhou 2014). Hierarchical relationships among labels are described in MIPS Functional Catalogue and Gene Ontology. Valentini (Valentini 2014) uses a binary classifier for each label according to True Path Rule (TPR) and the funCat. Recent work of Guoxian and the co-authors (Yu et al. 2015), explored the incomplete label problem in a hierarchical manner using function correlation. Another approach for predicting protein function, as proposed by Piovesan et al. (Piovesan et al. 2015) includes the combination of the trio: PPIN information, protein domain and sequence. In another work, Zhao et al. (Zhao et al. 2016) invokes dynamic weighted interaction network instead of the static one. This dynamic network is enriched with PPIN, time course gene expression data, protein’s domain information and protein complex information which ultimately predict function of a protein using majority ranking. While most of the predictive models highlights on the most highly related similar proteins in the neighborhood of the test protein, Reinders et al. (Reinders et al. 2018) focuses on the less similar proteins. It is shown by the application of Label-Space Dimensionality Reduction (LSDR) techniques that though these proteins are less similar but they are quite informative and plays an important role in protein function prediction. Another iterative algorithm is implemented by Sun et al. (Sun et al. 2018) for predicting protein functions. It is completely dependent on the identification of the functional dependencies which are based on proteins and their interactions. Sequence similarity network is another important aspect for protein function prediction, which is considered in the development of the methodology, Effusion, as proposed in the work of Yunes & Babbitt (Yunes & Babbitt 2018). Other notable works in this field are Wang et al. (Wang et al. 2018) and Fa et al. (Fa et al. 2018)

All these methods discussed above have already taken protein function prediction to the next higher level. Yet, the still uncovered details in the study and analysis support the need for new computational methods exploiting the protein-protein interaction networks for biological function identification. In our novel methodology FunPred 3.0, the functions of test proteins are determined by analyzing the neighbourhood properties of their protein interaction network. At the same time, certain selected physicochemical properties of amino acids are also used along with it. This task is challenging because of several reasons; e.g., large number of functional groups, different levels of the interconnection hierarchy, proteins with multiple functional groups, and incomplete or missing labels. In this proposed methodology, MIPS dataset (Mewes et al. 2002) is used. It contains protein pairs along with their corresponding functions. At the initial phase, to estimate the effectiveness of FunPred 3.0, essential proteins are selected as test proteins. The functions of these proteins are considered to be unknown for experimental purpose though their functions are defined in the dataset. Then we have applied FunPred 3.0 to predict the functions of the test proteins. Predicted functions are hence matched with the original ones to compute precision, recall and F-Score. While executing FunPred 3.0, it has been observed that 870 PPIs out of 13528 PPIs i.e., ~6.4% of the overall MIPS dataset (Mewes et al. 2002) are unpredicted i.e. either unknown or missing. FunPred 3.0 has been also applied to predict the unannotated protein function and protein interaction (Mamoon et al. 2010), and also assigning the functional annotations for 767 PPIs out of 870 PPIs, representing circa ~5.7% of the overall MIPS dataset. Similar instances have been also observed in the case of subcellular location of proteins where 1679 proteins out of 6721 unique number of proteins are still unpredicted i.e. the
subcellular localization of these proteins are still unknown. The predicted functional annotations and subcellular localization of these unpredicted proteins and protein pairs respectively result in relevant biological information, such as vital processes, diseases related mutations.

**METHODOLOGY**

In one of our two earlier works, Funpred-1 (Saha et al. 2014), the selection of ten percent of test proteins of the top eight functional groups from the dataset was done randomly. Top eight functional groups were selected on the basis of maximum number of occurrences and interactions of proteins in them. While in another, FunPred 2 (Saha et al. 2017a), protein clusters are formed initially by the application of node and edge weight. Then fifty percent of proteins from each of the formed clusters are selected as test proteins. In both the cases, test proteins are chosen randomly. Since both these works are completely based on protein function prediction from PPIN, so network formed for each test protein is extensively large (up to level – 2) enough to process which gradually enhances the overall computational overhead. So randomness is basically implemented to filter out the most essential proteins out of the entire PPIN and select them as the test protein. But, variation of test set (i.e. proteins beyond ten percent in FunPred-1 or fifty percent in FunPred 2) as well as application of node and edge weight thresholds (in FunPred 2) might also play an important role in prediction accuracy level which has not been yet tested so far. It may be also considered as a major drawback and limitation of our previous two methodologies, which has been tried to overcome in FunPred 3.0. FunPred 3.0 is an extended and advanced version of FunPred 2. Though the basic outlay of the both is same but the uniqueness of FunPred 3.0 can be defined in a three way approach:

1. Application of three levels of threshold for the formation of clusters and selecting all proteins from the clusters as test set to overcome the limitations of its predecessors.
2. Incorporation of feature selection.
3. Capable of prediction of functions of unpredicted protein pairs (incomplete and missing functional annotations) in MIPS dataset of *Saccharomyces cerevisiae*.
4. Capable of prediction of subcellular localization of proteins.

PPIN formed of proteins and their corresponding interactions may contain essential/non-essential proteins as well as reliable/unreliable edges. Proteins having maximum number of interconnected neighbors are considered as essential proteins while proteins having less number of interconnected neighbors are considered as non-essential ones. Presence of non-essential neighbors in the PPIN might affect the unknown protein function prediction level accuracy. So proper identification and elimination of non-essential proteins is needed to ensure the presence of maximum number of essential proteins in the PPIN. In the proposed work, detection of essential proteins is implemented by node weight. Node weight (Wang & Wu 2013) basically assigns a weightage score to each node or protein based on its corresponding degree. High node weight determines essential while low node weight detects non-essential protein. Thus non-essential proteins are discarded from PPIN along their corresponding edges. Even after this initial phase of PPIN refinement, there are still some unreliable edges present in the network. Since protein clusters are formed, so it is obvious that two nodes with an edge between them belong to the same cluster if they have high similarity. Edge between two nodes of high similarity is
considered as reliable edge while that of low similarity is denoted as unreliable edge. In the proposed work, detection of reliable edges is executed by edge weight. Edge weight (Wang & Wu 2013) also assigns a weightage score to each edge connected by two proteins in the terminals. Assignment of edge weight to an edge depends on the number of common neighbors between the two terminal proteins of the corresponding edge. More number of common neighbors signifies high similarity which in turn detects reliable edges. On the other hand, unreliable edges have low similarity since they have less number of common neighbors. Thus, unreliable edges are identified and pruned from the PPIN. Filtered out PPIN after these refinements, contains only essential proteins and reliable edges, which ultimately helps in enhancing prediction accuracy level.

In newly proposed algorithm, FunPred 3.0, first detects protein cluster and then selects all the proteins as test proteins from different predicted clusters. We have adopted the approach of forming protein cluster as mentioned in the work of Wang and Wu (Wang & Wu 2013). Protein clusters, thus formed, comprises of proteins belonging to any functional group. It results in accumulating larger number of functional groups as compared with only 8 functional groups in our previous work (Saha et al. 2014). The novel computational method works in two stages:

1. All the unique proteins are first clustered into M mutually exclusive clusters based on their node weight and edge weight in the overall PPIN. Node and edge weight have been used to ensure all the most essential nodes with higher reliability are present in the cluster and get selected as test proteins.

2. Functional annotations are then derived from the multi-level neighbourhood of an unknown protein within each cluster.

More specifically, FunPred 3.0 is categorized into two sections: FunPred 3.0_Clust and FunPred 3.0_Pred.

FunPred 3.0_Clust uses the node weight and edge weight properties to rank and cluster all the proteins, creating M mutually exclusive protein clusters (Wang & Wu 2013). The number of functional labels, associated with each interacting pair is large and in some cases annotations in each such cluster is unpredicted (incomplete or missing). This fact forces us to heuristically choose the node and edge weight threshold values, such that the unlabelled proteins are associated with larger protein clusters and have many neighbourhood interactions (see Figure 1 and Figure 2). Three thresholds (high, medium and low) are set for each of node and edge weight using equation 1 (Zhang et al. 2016).

\[ Th_k = \alpha + k.\sigma.(1 - \frac{1}{1+\sigma^2}) \]  

(1)

where for node weight/edge weight, \( k \in \{1,2,3\} \) denotes three different thresholds i.e. low, medium and high respectively. \( \alpha \) is the mean of node weight/edge weight values of all proteins. \( \sigma \) is the standard deviation of node weight/edge weight values of all proteins. Proteins and edges having value less value than these node and edge weight thresholds get discarded and are considered as non-essential proteins and unreliable edges in the network respectively.

The entire methodology has been described in Algorithm 1 as well as pictorially highlighted in Figure 1 and Figure 2. In Figure 1, sample Table 1 (node weight table) is formed.
from the initial PPIN of yeast. Hence, node weight threshold is calculated using equation 1 at
three levels: high, medium and low. These thresholds are applied on the initial PPIN to filter out
three sub-networks i.e. sub-network1, sub-network2, sub-network3 respectively at high, medium
and low node weight thresholds. Respective edge weight tables i.e. sample Table 2 (in Figure 1),
sample Table 3 (in Figure 1), sample Table 4 (in Figure 1) are formed from sub-network1, sub-
network2, sub-network3, upon which high, medium and low edge weight thresholds (obtained
using equation 1) are applied to form pruned sub-network4, sub-network5 and sub-network3.

In Figure 2, sample Table 5 (node weight table formed from sample Table 1 under high
threshold in Figure 1) has been generated from refined sub-network4 which afterwards is sorted
in descending order to form sample Table 6 (in Figure 2). From sample Table 6 (in Figure 2),
first protein having the highest node weight is selected as seed of the initial cluster. Then
corresponding level 1 neighbors of the seed are included in the cluster provided its inclusion in
the cluster does not let the edge weight to fall below the high threshold (verified using sample
Table 2 in Figure 1). The entire initial cluster contents are discarded from sample Table 6 (in
Figure 2) and the next cluster is formed with the next selected seed. This process continues until
all the proteins in sample Table 6 (in Figure 2) get clustered. Thus M mutually exclusive protein
clusters are formed where 1 =< M <= N4 (N4 is the total number of proteins present in sample
Table 6 in Figure 2). The entire procedure is repeated for refined sub-network5 and refined sub-
network6 until M mutually exclusive protein clusters is formed for each of them.

All the proteins belonging to M mutually exclusive protein clusters (under three levels of
thresholds: high, medium and low separately), obtained from FunPred 3.0_Clust are considered
as essential proteins and hence they are included in the test set of our proposed methodology. In
the second step, FunPred 3.0_Pred predicts labels these selected unlabelled (or test) proteins
using neighbourhood properties and physicochemical properties of amino acids (see Algorithm 2
and Figure 3). In Figure 3, for each test protein (say P1 belonging to refined sub-network4 in
Figure 1) its level 1 neighborhood graph is formed including those proteins which are present in
its corresponding node weight table: sample Table 1 under high threshold in Figure 1. Then
protein clusters are formed at level 1 (considering each level 1 protein as the seed of the cluster)
in a similar way as formed in FunPred 3.0_Clust. It should be noted here that the number of
clusters formed here is equivalent to the number of proteins in level 1. Distance between the
mean of the physico-chemical features of each protein cluster as well as test protein is computed
and the test belongs to the cluster having the least distance. All the functions of the selected
cluster are allocated to the test protein.

The relevant level – 1 neighbours of the test proteins are chosen to form their individual
neighbourhood graph. In finding the level – 1 neighbours or forming their individual
neighbourhood graph, relevance is measured in terms of edge weight properties. Next, PCP
score is computed for the neighbourhood graph of each test protein. Six different high-ranked
physico-chemical features: aliphatic index (Singh et al. 2008), gravy (Kyte & Doolittle 1982),
aromacity (Lobry & Gautier 1994), number of negatively charged residues (Singh et al. 2008),
number of positively charged residues (Singh et al. 2008), isoelectric point (Bjellqvist et al.
1994) are used to reckon this physico-chemical property (PCP) based score. These high-ranked
features are selected from ten divergent physico-chemical features (see supplementary) by the
enactment of four distinct classifiers: XGBoost classifier (Chen & Guestrin 2016; Pedregosa et
al. 2011), Random Forest classifier (Breiman 2001; Pedregosa et al. 2011), Extra Tree classifier
(Geurts et al. 2006; Pedregosa et al. 2011) and Recursive feature elimination classifier (Pedregosa et al. 2011). High-ranked five among ten features have been picked at first by each classifier. Then from these picked features, frequency of maximum occurrences for each individual feature has been noted from which endmost six features get selected (see Table 1). Finally, each test protein is assigned to a functional group of the neighbourhood graph, based on the nearest neighbourhood approach on the basis of mean PCP score. FunPred 3.0_Clust (see Algorithm 1) and FunPred 3.0_Pred (see Algorithm 2) describes the methodology of unknown protein selection and function prediction respectively.
Algorithm 1. FunPred 3.0_Clust:
(For formation of protein clusters which consist of essential proteins and reliable edges)

Input: Undirected PPIN \( G \)
Output: Protein clusters at three levels of threshold: high, medium and low

Begin

//computation of node weight of \( G \)
for all nodes in \( G \)
compute node weight

//computation of node weight threshold
compute node weight threshold at three levels: high, medium and low using equation 1

//Elimination of non-essential proteins based on node weight threshold
for each level of threshold
for all nodes in \( G \)
if node weight does not exceed threshold
remove corresponding node.

//Formation of refined sub-networks \( G_{\text{high}} \), \( G_{\text{medium}} \) and \( G_{\text{low}} \) from \( G \)
\( G_{\text{high}}, \ G_{\text{medium}} \) and \( G_{\text{low}} \) consisting of only essential proteins (high node weight) are formed

//computation of edge weight of \( G \)
for all edges in \( G_{\text{high}}, G_{\text{medium}} \) and \( G_{\text{low}} \)
compute edge weight

//computation of edge weight threshold
compute edge weight threshold at three levels: high, medium and low using equation 1

//Elimination of unreliable edges based on edge weight threshold
for all edges in \( G_{\text{high}} \)
if edge weight does not exceed high level of edge threshold
remove corresponding edge.
repeat the same for low, medium level of threshold and \( G_{\text{medium}} \) and \( G_{\text{low}} \) respectively.

//Formation of refined sub-networks \( G_{\text{high,high}}, G_{\text{medium,medium}} \) and \( G_{\text{low,low}} \) from \( G_{\text{high}}, G_{\text{medium}} \) and \( G_{\text{low}} \) at high node and edge weight threshold, medium node and edge weight threshold, low node and edge weight threshold respectively.
form \( G_{\text{high,high}}, G_{\text{medium,medium}} \) and \( G_{\text{low,low}} \) consisting of only reliable edges (high edge weight)

//Formation of clusters at three levels of thresholds
for all proteins in \( G_{\text{high,high}} \)
form node weight table
sort the node weight table based on the node weights
select the first protein \( P \) in the node weight table as the seed of initial cluster \( C_{M}^{\text{high}} \)
i.e. \( C_{M}^{\text{high}} = \{ P \} \) where \( 1 \leq M \leq W \) (W is the total no. of nodes in node weight table)
nearighbors of \( P \) are added to \( C_{M}^{\text{high}} \) provided its inclusion does not cause edge weight to fall below high edge weight threshold value i.e. \( C_{M}^{\text{high}} = \{ P \} \cup N_{P} \)
update the node edge table by eliminating all the proteins present in \( C_{M}^{\text{high}} \) and continue with the next seed to form clusters in the same way mentioned above till all the proteins in the node weight table belongs to a cluster.
repeat the same procedure for \( G_{\text{medium,medium}} \) and \( G_{\text{low,low}} \).
**Algorithm 2.** FunPred 3.0_Pred:
(Protein function prediction of test proteins)

| Input: | Set of un-annotated proteins in $C^{high}_M$, $C^{medium}_M$, $C^{low}_M$ selected by FunPred 3.0_Clust |
|-------|--------------------------------------------------------------------------------------------------|
| Output: | Functional group of un-annotated proteins |

Begin

// Formation of clusters at level – 1 of un-annotated protein
for each protein $P$ in $C^{high}_M$
    for each level – 1 neighbor $N$ of $P$ present in $C^{high}_M$
        add $N$ as the seed of the cluster $K_i$, i.e. $K_i = \{N\}$.
//where $I = i <= g$, $g$ is the total number of level – 1 neighbors of $P$
        add immediate neighbors of $N$ i.e. $IN_N$ in the cluster $K_i$ provided such
        inclusion does not cause the edge weight to fall below the high edge weight
        value of threshold as computed in FunPred 3.0_Clust i.e. $K_i = \{N\} \cup IN_N$.

// Feature selection
compute Physico-Chemical features of each protein from the amino acid sequence
of each protein and execute the selected classifiers to select the most essential
features.
// Here six features get selected as the essential ones from the initial list of ten
features.

// Computation of $PCP_{score}$
for each protein $P$ in $C^{high}_M$
    compute its mean $PCP_{score}$ of six selected Physico-Chemical features
    for each formed cluster $K_i$ of protein $P$
        compute its mean $PCP_{score}$ of six selected Physico-Chemical features

// Assigning of Functional Groups to the proteins in $C^{high}_M$
for each protein $P$ in $C^{high}_M$
    for all clusters $K_i$ of protein $P$
        obtain the difference of $PCP_{score}$ of $P$ and clusters $K_i$
        functional groups of cluster $K_i$ are assigned to protein $P$ having least difference.
    Repeat all the above steps for annotation of protein functions in $C^{medium}_M$ and $C^{low}_M$

End.
It needs to be highlighted here that both FunPred 3.0_Clust and FunPred 3.0_Pred have been executed at three levels: 1. High node and edge weight threshold, 2. Medium node and edge weight threshold, 3. Low node and edge weight threshold. So, we have tested FunPred 3.0 at each of the three levels to assess its performance impact.

Besides, predicting protein function, it has been observed that the protein subcellular localization is yet another important aspect which needs to be considered since it helps in better understanding of protein function. So, subcellular localization dataset of yeast has been obtained from UniProt database (Apweiler et al. 2004). On careful observation it has been noted that there are 6721 unique number of proteins out of which localization of some proteins are still unpredicted. It is similar as that of 6.4% of the overall MIPS dataset (Mewes et al. 2002) which are unpredicted i.e. either unknown or missing. So FunPred 3.0 is also implemented to predict these unpredicted protein subcellular localization.

In FunPred 3.0, the subcellular localization dataset of yeast is centrally categorized under three major sections: proteins residing in nucleus (termed as nuclear proteins), proteins residing in cytoplasm (termed as cytoplasm proteins) and proteins residing in other regions (termed as interface proteins). Besides these, there is also another section termed as unpredicted localization proteins, consisting of those whose localization are not yet predicted (see Figure 4, Figure 5 and Figure 6). So before dealing with the unpredicted localization proteins, the predictive accuracy of FunPred 3.0 needs to be assessed just in a similar way earlier as that of protein function prediction. Same test set of essential proteins (as generated by FunPred 3.0_Clust) of yeast (MIPS) is also considered here (localization of which are known but considered to be unknown for experimental purpose). Selected test and candidate proteins in the PPIN of nuclear, cytoplasm and interface proteins are highlighted in Figure 7, Figure 8 and Figure 9 respectively. Now for each test protein, its corresponding level – 1 neighborhood graph is formed. In a PPIN, a protein almost shares similar properties as that of its neighborhood proteins. The same is also applicable to the test protein but all the properties or functions of the neighborhood cannot be transmitted to it. So proper assessment needs to be implemented in the neighborhood of the test protein. For this purpose, FunPred 3.0_Pred_SL (SL stands for Subcellular Localization) is applied. It first assigns respective subcellular localization information to the neighborhood of the test protein using UniProt database. Hence it counts the frequency of occurrence of each subcellular location (nucleus, cytoplasm or any other region). The subcellular location having the highest frequency of occurrence among the neighborhood is allocated to the test protein. The test protein becomes nuclear or cytoplasm or interface proteins according to the allocated subcellular location. Then the allocated subcellular localization is checked from the Uniprot database. The overall result which is achieved by the application of FunPred 3.0_Pred_SL, has been highlighted in Table 2. An overall accuracy of 69.1%, 57% and 53% is reckoned for nuclear, cytoplasm and interface proteins respectively. It is observed from Table 2, that our method fails to predict the subcellular localization of few proteins among all the three categories. This is because our method mainly predicts subcellular localization using PPIN neighborhood based approaches and if there is practically no information or significantly less amount of interactive information in the PPIN for a particular test protein then our method fails. This can be considered as one of our limitations which can be redressed in future works by incorporating protein sequence.

Subcellular localization dataset of yeast contains 1679 number of unpredicted localization proteins out of which our method predicts the localization of 638 proteins successfully.
Localization information for the remaining 1041 proteins cannot be predicted because of the absence of PPIN interaction in MIPS dataset as discussed earlier. This extra added layer of biological information about subcellular localization of proteins along with the protein function prediction boost up our methodology FunPred 3.0 to the next higher level.

**RESULTS**

Initially, PPIN of yeast consists of 4554 unique proteins in 13528 protein-protein interactions (PPINs) after the elimination of the self-replicating and the self-interacting protein pairs. After the network refinement through the execution of node and edge weight threshold, non-essential proteins along with unreliable edges get eliminated and the initial PPIN gets reduced to almost 3174 unique proteins and 6936 PPINs (approx.) considering three levels of thresholds from which FunPred 3.0_Clust form protein clusters to generate test set of proteins.

During the result analysis it is observed that proteins belonging to random functional groups like lipid metabolism, DNA Repair etc. get selected as test proteins. In FunPred 3.0_Pred, all proteins from each protein cluster formed from FunPred 3.0_Clust are considered as test proteins. The overall initial PPIN of yeast is highlighted in Figure 10 while in Figure 11, 433 test proteins (most essential ones among the refined PPIN of yeast consisting of 3174 unique proteins and 6936 PPINs) selected by FunPred 3.0_Clust at high node and edge weight threshold values, are highlighted in yellow circle (shape) in the initial PPIN of yeast consisting of 4554 unique proteins in 13528 protein-protein interactions (PPINs). It should be noted in Figure 11, that all the selected test proteins belongs to the most densely connected region of PPIN which establishes the fact that these are indeed most strongly connected essential proteins. The performance of FunPred 3.0 is evaluated using standard performance measures, such as Precision (P), Recall (R) and F-Score (F) values, which are calculated using the following equations:

\[
P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \quad F = \frac{2 \times (P \times R)}{P + R}
\]

where TP, FP, FN represent True Positive, False Positive and False Negative respectively.

The performance of FunPred 3.0 has been analyzed under different levels of thresholds of node and edge weights as highlighted in Table 3. It should be noted here that under high and medium thresholds, same Precision, Recall and F-score have been retrieved since number of selected test proteins are equivalent in both the cases. All over result analysis as depicted in Table 3 shows that there are not much significant changes in the result as such varying the thresholds except a slight fall in precision and F-score under low threshold as compared to the others. The fall is due to the relaxation in the node and edge weight thresholds resulting in incorporation of less essential proteins in the test set. So, the high threshold ensures inclusion of the most essential proteins. So the Precision, Recall and F-score of FunPred 3.0 are reckoned as 0.55, 0.82 and 0.66 respectively under high threshold. High Recall and low Precision emerges out as a major characteristic of FunPred 3.0 when compared to the other methodologies except FunPred-2. It highlights the fact that most of the admissible results are successfully generated by FunPred 3.0. Table 4 shows a detailed performance comparison of other methodologies along with our proposed systems (like FunPred 1.1, FunPred 1.2 (Saha et al. 2014)), neighbourhood 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), FPred_Apriori (Prasad et al. 2017), Zhang methodology (Zhang et al. 2009), Domain Combination Similarity (DCS) (Peng et al. 2014), Domain Combination Similarity in context of protein complexes (DSCP) (Peng et al. 2014), Protein Overlap Network (PON) (Liang et al. 2013), Deep_GO (Kulmanov et al. 2018) and the FS-weight based method (Chua et al. 2006)). All these data are collected from their respective works which are executed on the same organism i.e. yeast. The results of Deep_GO (Kulmanov et al. 2018) are computed manually for yeast dataset, the code of which is available at https://github.com/SovanSaha/FunPred-3.0.git. From Table 4, it can be also highlighted that our method, FunPred 3.0, yields relatively higher F-Score values than the others including its earlier version FunPred-2.

Table 4 also discloses the fact that NRC method has overpowered the rest except FunPred 1.1 (Saha et al. 2014), FunPred 1.2 (Saha et al. 2014), FPred_Apriori (Prasad et al. 2017) and FunPred 3.0. The reason behind this is observed as follows: Both version of FunPred 1 has incorporated two levels (i.e. level – 1 and level – 2) of PPIN as well as lot of essential neighborhood properties like neighborhood ratio, protein path connectivity and relative functional similarity (includes both ancestor and descendant information of a specific protein) have been utilized to assess the reliability of each node (protein) along with its associated edges (protein interaction) during the unannotated protein function prediction. FPred_Apriori (Prasad et al. 2017) executes both closeness centrality and edge clustering coefficient to make its predictive approach more effective than the others. Last but not least, FunPred 3.0 combines physico-chemical property of each protein along with neighborhood analysis (like node weight, edge weight etc.) for predicting protein function which ultimately promotes it to the next higher level in the terms of performance analysis when compared to the others.

Though Neighborhood counting method is simple in nature yet the performance measure of it has descended considerably in comparison to NRC (Moosavi et al. 2013), FS-weight #1 (directly connected proteins) and FS-weight #1 and #2 (directly and indirectly connected proteins) despite of its simplicity (Chua et al. 2006). This is because no differentiation has been observed between the direct and indirect neighborhood connection. Beside most of the methods included in Table 4 like NRC (Moosavi et al. 2013), Chi square #1&2 (Hishigaki et al. 2001), Chi square #1 (Hishigaki et al. 2001), Neighborhood counting #1&2 (Schwikowski et al. 2000), Neighborhood counting #1 (Schwikowski et al. 2000) etc. are not utilized for the refinement of the PIN by pruning unreliable proteins or edges which in turn increases false positives in their prediction accuracy level. In FPred_Apriori (Prasad et al. 2017), a bottom-up predictor of existing Apriori algorithm has been utilized for protein function prediction by exploiting two most important neighborhood properties: closeness centrality and edge clustering coefficient of protein interaction network. Though the method is unique in the fact that the functions of the leaf nodes in the interaction network have been back propagated and thus labeled up to the root node (test protein) but yet it fails to generate high Recall and F-score than FunPred 3.0. But it returns substantially high precision values than the others as well as all our methods. DCS (Peng et al. 2014), DSCP (Peng et al. 2014), PON (Liang et al. 2013), Deep_GO (Kulmanov et al. 2018) and Zhang methodology (Zhang et al. 2009) are well developed methods for protein function prediction incorporating domain specific as well as neighborhood based properties but they fail to compete with all our methodologies due to the lack of important feature selection methodologies of physicochemical properties and proper assessment of nodes and edges involved in test set protein function prediction through node and edge weights.
During experimental evaluation, the validation set is prepared with 4554 labeled *Saccharomyces cerevisiae* proteins, collected from the MIPS dataset. Using FunPred 3.0_Clust, we identify M mutually exclusive protein clusters (Wang & Wu 2013). Experimental variations with \( k = 1,2,3 \) are included in Table 2. Using an optimal choice of \( k = 3 \), we identify 433 test targets for the validation set. Now, the functional labels of these test proteins are assigned using FunPred 3.0_Pred. The Precision, Recall, F-scores of our method over the test targets of the validation set is obtained as 0.55, 0.82 and 0.66 respectively.

**DISCUSSION**

Our results (characterized by the Precision, Recall and F-Score) and comparison with the other protein functional group prediction models show the superiority of our approach. The FunPred 3.0 software has better performance than any existing function prediction *in silico* method. The network structure may be pruned based on the edge weight and along with it use of physico-chemical properties lead to improved and faster functional prediction in complex and diverse protein–protein interaction networks. We would like to estimate the effectiveness of our in silico method for other organisms, such as in human protein–protein interactions with even more complex network architectures.

The initial results motivate us to predict the subcellular localization and unpredicted protein pair functions (missing/unknown functions) for 870 PPIs extracted from MIPS dataset. A protein can perform multiple functions in isolation. It may also perform some specific functions while interacting with one protein while perform some other specific functions while reacting with other proteins. But considering the fact that a protein often shares similar functions with proteins that interact with it (Chakicherla et al. 2018; Chatterjee et al. 2012; Shatsky et al. 2016), each protein pair is disintegrated in two constituent proteins and functions of each protein is predicted using FunPred 3.0. For an unknown protein pair \( P_1P_2 \), we predict the functions as an intersection of FunPred 3.0_Pred\( (P_1) \cap \)FunPred 3.0_Pred\( (P_2) \). The results of all the predicted annotations for the MIPS dataset are available at https://github.com/SovanSaha/FunPred-3.0.git. Examples of the prediction of protein function and interactions for unpredicted pairs (both unknown and missing protein pair) have been shown in Table 5 and Table 6 respectively.

Summarizing, 767 unpredicted protein pair functions (511 unknown protein pair functions and 256 missing protein pair functions) in the MIPS dataset could be predicted using our FunPred 3.0 algorithm. Our approach failed to predict 103 unpredicted protein pairs since they have less number of acceptable neighbors. Simultaneously our methodology also performs very well in predicting subcellular localization of proteins as discussed earlier in the methodology section earlier. All the datasets and supplementary files are also freely available at https://github.com/SovanSaha/FunPred-3.0.git.

**CONCLUSION**

FunPred 3.0 thus proved to be an improved and advanced version of our previous
methodology FunPred-2. The enhanced performance of FunPred 3.0 is due to the use of node weight, edge weight, and physicochemical properties of proteins in the prediction pathway of test set of proteins. It should be highlighted here that FunPred 3.0 incorporates the most essential features classified through four classifiers: XGBoost, Random Forest, Extra Tree and Recursive feature elimination. Recursive feature elimination (RFE) which indeed plays an important role in improving the performance of the proposed methodology. Though this method does not consider dynamic PPIN and integration of other multiple types of data like domain (Chatterjee et al. 2011a; Chatterjee et al. 2011b) etc., but topological analysis, association between function and protein have been proven to be significant for this research. Besides use of FunPred 3.0 to detect the subcellular localization of proteins as well as function of unpredicted protein pair functions (unknown and missing pairs of proteins) in MIPS database add an extra dimension to this work. Incorporation of other protein related features and their integration, use of the other benchmark datasets for different organism may give a proper insight of prediction. Beside this unannotated protein function prediction, the methodology behind the FunPred 3.0 algorithm can be also used in disease specific datasets (Saha et al. 2017b) also which may be a future direction as well. In a nutshell, the work presented here proposes the statistical learning evaluation of various features for prediction of protein functions in the complex yeast PPIN with reasonable accuracy. The dataset used in this study and the complete source codes of the FunPred 3.0 software package are available in the public domain (https://github.com/SovanSaha/FunPred-3.0.git) for non-commercial research.

REFERENCES

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O’Donovan C, Redaschi N, and Yeh L-SL. 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Research 32(Database issue):D115-D119. 10.1093/nar/gkh131.

Bjellqvist B, Basse B, Olsen E, and Celis JE. 1994. Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. ELECTROPHORESIS 15(1):529-539. 10.1002/elps.1150150171.

Breiman L. 2001. Random Forests. Machine Learning 45(1):5-32. 10.1023/a:1010933404324.

Chakicherla A, Ligon Dang M, Rodriguez V, Hansen J, and Zemla A. 2018. Function prediction of an interacting protein pair using domain fusion analysis: SpaR and SpaK.

Chatterjee P, Basu S, Kundu M, Nasipuri M, and Plewczynski D. 2011a. PPI_SVM: Prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables. Cellular & Molecular Biology Letters 16(2):264-278. 10.2478/s11658-011-0008-x.

Chatterjee P, Basu S, Kundu M, Nasipuri M, and Plewczynski D. 2011b. PSP_MCSVM: brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machines. Journal of Molecular Modeling 17(9):2191. 10.1007/s00894-011-1102-8.

Chatterjee T, Chatterjee P, Basu S, Kundu M, and Nasipuri M. 2012. Protein function by minimum distance classifier from protein interaction network. In: 2012 International Conference on Communications, Devices and Intelligent Systems (CODIS).588-591.

Chen J, Hsu W, Lee ML, and Ng SK. 2007. Labeling network motifs in protein interactomes for protein function prediction. In: 2007 IEEE 23rd International Conference on Data Engineering.546-555.
Chen T, and Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: ACM.785-794.

Chua HN, 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. 10.1093/bioinformatics/btl145.

Fa R, Cozzetto D, Wan C, and Jones DT. 2018. Predicting human protein function with multi-task deep neural networks. PLOS ONE 13(6):e0198216. 10.1371/journal.pone.0198216.

Geurts P, Ernst D, and Wehenkel L. 2006. Extremely randomized trees. Machine Learning 63(1):3-42. 10.1007/s10994-006-6226-1.

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. 10.1002/yea.706.

Jiang JQ, and McQuay LJ. 2012. Predicting Protein Function by Multi-Label Correlated Semi-Supervised Learning. IEEE/ACM Transactions on Computational Biology and Bioinformatics 9(4):1059-1069. 10.1109/TCBB.2011.156.

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

Kyte J, and Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology 157(1):105-132. https://doi.org/10.1016/0022-2836(82)90515-0.

Lee D, Redfern O, and Orengo C. 2007. Predicting protein function from sequence and structure. Nature Reviews Molecular Cell Biology 8(995. 10.1038/nrm2281

Liang S, Zheng D, Standley DM, Guo H, and Zhang C. 2013. A novel function prediction approach using protein overlap networks. BMC Systems Biology 7(1):61. 10.1186/1752-0509-7-61.

Lichtarge O, Bourne HR, and Cohen FE. 1996. An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families. Journal of Molecular Biology 257(2):342-358. https://doi.org/10.1006/jmbi.1996.0167.

Lobry JR, and Gautier C. 1994. Hydrophobicity, expressivity and aromaticicity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Research 22(15):3174-3180.

Mamoon R, Sumathy R, and Gajendra PSR. 2010. A Simple Approach for Predicting Protein-Protein Interactions. Current Protein & Peptide Science 11(7):589-600. http://dx.doi.org/10.2174/138920310794109120.

Mewes HW, Frishman D, Güldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Münsterkötter M, Rudd S, and Weil B. 2002. MIPS: a database for genomes and protein sequences. Nucleic Acids Research 22(15):3174-3180.

Mills CL, Beuning PJ, and Ondrechen MJ. 2015. Biochemical functional predictions for protein structures of unknown or uncertain function. Computational and Structural Biotechnology Journal 13(182-191. https://doi.org/10.1016/j.csbj.2015.02.003.

Moosavi S, Rahgozar M, and Rahimi A. 2013. Protein function prediction using neighbor relativity in protein–protein interaction network. Computational Biology and Chemistry 43(11-16. https://doi.org/10.1016/j.compbiolchem.2012.12.003.
Ng PC, and Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* 31(13):3812-3814.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, and Duchesnay E. 2011. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 12(2825-2830).

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 Systems Biology* 8(1):35. 10.1186/1752-0509-8-35.

Piovesan D, Giollo M, Leonardi E, Ferrari C, and Tosatto SCE. 2015. INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity. *Nucleic Acids Research* 43(Web Server issue):W134-W140. 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. In: Computational Intelligence, Communications, and Business Analytics. Singapore: Springer Singapore.3-16.

Reinders MJT, van Ham RCHJ, and Makrodimitris S. 2018. Improving protein function prediction using protein sequence and GO-term similarities. 10.1093/bioinformatics/bty751.

Saha S, Chatterjee P, Basu S, Kundu M, and Nasipuri M. 2012. Improving prediction of protein function from protein interaction network using intelligent neighborhood approach. In: 2012 International Conference on Communications, Devices and Intelligent Systems (CODIS).584-587.

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. *Cellular & Molecular Biology Letters* 19(4):675-691. 10.2478/s11658-014-0221-5.

Saha S, Chatterjee P, Basu S, and Nasipuri M. 2017a. Functional Group Prediction of Un-annotated Protein by Exploiting Its Neighborhood Analysis in Saccharomyces Cerevisiae Protein Interaction Network. In: Chaki R, Saeed K, Cortesi A, and Chaki N, eds. *Advanced Computing and Systems for Security: Volume Four*. Singapore: Springer Singapore, 165-177.

Saha S, Sengupta K, Chatterjee P, Basu S, and Nasipuri M. 2017b. Analysis of protein targets in pathogen–host interaction in infectious diseases: a case study on Plasmodium falciparum and Homo sapiens interaction network. *Briefings in Functional Genomics*:elix024-elix024. 10.1093/bfgp/elix024.

Schwikowski B, Uetz P, and Fields S. 2000. A network of protein–protein interactions in yeast. *Nature Biotechnology* 18(1257. 10.1038/82360

Singh M, Wadhwa PK, and Kaur S. 2008. Predicting Protein Function using Decision Tree. *World Academy of Science, Engineering and Technology* 2(3):300-303.

Sriwastava BK, Basu S, and Maulik U. 2015. Predicting Protein-Protein Interaction Sites with a Novel Membership Based Fuzzy SVM Classifier. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 12(6):1394-1404. 10.1109/TCBB.2015.2401018.
Sun P, Tan X, Guo S, Zhang J, Sun B, Du N, Wang H, and Sun H. 2018. Protein Function Prediction Using Function Associations in Protein–Protein Interaction Network. *IEEE Access* 6(30892-30902). 10.1109/ACCESS.2018.2806478.

Valentini G. 2014. Hierarchical Ensemble Methods for Protein Function Prediction. *ISRN Bioinformatics* 2014(34). 10.1155/2014/901419.

Vazquez A, Flammini A, Maritan A, and Vespignani A. 2003. Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology* 21(697). 10.1038/nbt825.

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

Wang Z, Zhao C, Wang Y, Sun Z, and Wang N. 2018. PANDA: Protein function prediction using domain architecture and affinity propagation. *Scientific Reports* 8(1):3484. 10.1038/s41598-018-21849-1.

Xiong W, Liu H, Guan J, and Zhou S. 2013. Protein function prediction by collective classification with explicit and implicit edges in protein-protein interaction networks. *BMC Bioinformatics* 14(Suppl 12):S4-S4. 10.1186/1471-2105-14-S12-S4.

Yu G, Zhu H, and Domeniconi C. 2015. Predicting protein functions using incomplete hierarchical labels. *BMC Bioinformatics* 16(1). 10.1186/s12859-014-0430-y.

Yunes JM, and Babbitt PC. 2018. Effusion: prediction of protein function from sequence similarity networks. *Bioinformatics* 35(3):442-451. 10.1093/bioinformatics/bty672.

Zhang ML, and Zhou ZH. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26(8):1819-1837. 10.1109/TKDE.2013.39.

Zhang S, Chen H, Liu K, and Sun Z. 2009. Inferring protein function by domain context similarities in protein-protein interaction networks. *BMC Bioinformatics* 10(1):395. 10.1186/1471-2105-10-395.

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 Bioinformatics* 17(7):229. 10.1186/s12859-016-1101-y.

Zhao B, Wang J, Li M, Li X, Li Y, Wu FX, and Pan Y. 2016. A New Method for Predicting Protein Functions From Dynamic Weighted Interactome Networks. *IEEE Transactions on NanoBioscience* 15(2):131-139. 10.1109/TNB.2016.2536161.
Figure 1

Application of Node Weight and Edge Weight at three levels of threshold: High, Medium and Low in FunPred 3.0_Clust

*Application of Node Weight threshold: Nodes having Node Weight below Node Weight Threshold are discarded.

*Application of Edge Weight threshold: Edges having Edge Weight below Edge Weight Threshold are discarded.
Figure 2

Formation of clusters from refined network after application of three levels of Node and Edge Weight threshold in FunPred 3.0_Clust

1. Form Sample Table 5 from Refined Sub-Network4.
2. Form Sample Table 6 from Sample Table 5 by sorting Node Weight in descending order from Refined Sub-Network4.
3. Select $p_i$ as the seed of the initial cluster and include its level 1 proteins which are present in Sample Table 1 in Figure 1 under high threshold.
4. Include level 1 neighbors (edges) of the seed from Sample Table 2 in Figure 1 in the cluster of $p_i$.
5. Delete all proteins in the cluster of $p_i$ from Sample Table 6 and repeat the previous steps for the next highest Node Weight (protein) as seed for the next cluster till all the proteins in Sample Table 6 belongs to a cluster. Thus M-clusters are formed for Refined Sub-Network4.
6. Similar M-clusters are also formed for Refined Sub-Network5 and Refined Sub-Network6.
Figure 3

Working Model of FunPred 3.0

Formation of PPI network of P1

1. Mean of the Physico-chemical property score of all proteins in each cluster is computed and considered as the mean of the respective cluster.
2. Similarly mean of Physico-chemical property score of the unannotated protein is also computed.
3. Test protein is assigned to the cluster along with it’s functions having the least mean difference of the computed score.

Cluster C1

Test protein P1 belongs to cluster C1 along with all it’s functions

Predicted cluster for test protein
Figure 4

PPIN of Yeast (*Saccharomyces cerevisiae*): Cytoplasm proteins (red), Nuclear Proteins (Green), Interface Proteins (Blue), Unpredicted localization Proteins (Orange)
Figure 5

Sequential formation of Cytoplasm proteins (red), Nuclear Proteins (Green), Interface Proteins (Blue), Unpredicted localization Proteins (Orange) in PPIN of yeast
Figure 6

Separate PPIN's of Cytoplasm proteins (red), Nuclear Proteins (Green), Interface Proteins (Blue), Unpredicted localization Proteins (Orange) and their interactions
Figure 7

Candidate (green) and test (yellow) proteins in Nuclear PPIN (green and yellow) of Yeast (violet: Other nodes in the network)
Figure 8

Candidate (red) and test (yellow) proteins in Cytoplasm PPIN (red and yellow) of Yeast (violet: Other nodes in the network)
Figure 9

Candidate (blue) and test (yellow) proteins in Interface PPIN (blue and yellow) of Yeast (violet: Other nodes in the network)
Figure 10

PPI network of Yeast (*Saccharomyces cerevisiae*)
Figure 11

PPI network of annotated (red circle) and test/unannotated proteins (yellow circle) of Yeast network (*Saccharomyces cerevisiae*)
Table 1 (on next page)

Top-ranked selected physicochemical features (marked in blue) using 4 classifiers based on the maximum number of hits
| Physicochemical properties | Classifiers Used | Random Tree | Extra Tree | Recursive Feature Elimination | #Hits |
|---------------------------|------------------|-------------|------------|-------------------------------|-------|
|                           | XGBoost          | ✓           | ✓          |                               |       |
| Aromacity                 | ✗                | ✓           | ✓          | ✗                             | 2     |
| Gravy                     | ✗                | ✓           | ✓          | ✓                             | 3     |
| Instability index         | ✗                | ✓           | ✗          | ✗                             | 1     |
| Isoelectric point         | ✓                | ✓           | ✓          | ✓                             | 4     |
| Negatively charged particle | ✓      | ✗           | ✗          | ✓                             | 2     |
| Positively charged particle | ✓   | ✗           | ✓          | ✓                             | 3     |
| Extinction coefficient    | ✓                | ✗           | ✗          | ✗                             | 1     |
| Aliphatic index           | ✓                | ✓           | ✓          | ✗                             | 3     |
| Absorbance                | ✗                | ✗           | ✗          | ✓                             | 1     |
| Ip/mol weight             | ✗                | ✗           | ✗          | ✓                             | 0     |
Table 2 (on next page)

Performance analyses of FunPred 3.0_Pred_SL
| Types of Proteins (based on Subcellular-localization) | Total no. of proteins in database | Total number of selected annotated proteins | Total number of selected essential test proteins | Prediction Accuracy (Total no. of matched proteins) | Prediction Accuracy (Total no. of Unmatched proteins) | Failed to predict |
|----------------------------------------------------|----------------------------------|---------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|------------------|
| Nuclear Proteins                                    | 1771                            | 1609                                        | 162                                             | 112                                             | 32                                             | 18               |
| Cytoplasm Proteins                                  | 1757                            | 1566                                        | 191                                             | 109                                             | 51                                             | 31               |
| Interface Proteins                                  | 2246                            | 2176                                        | 70                                              | 37                                              | 23                                             | 10               |
Table 3 (on next page)

Precision, Recall and F-score obtained at three levels of node and edge weight threshold
| Threshold Type | Node Weight Threshold | Edge Weight Threshold | Selected Test Proteins | Precision | Recall | F-score |
|----------------|-----------------------|-----------------------|------------------------|-----------|--------|---------|
| High           | 1.072                 | 0.110                 | 433                    | 0.55      | 0.82   | 0.66    |
| Medium         | 1.068                 | 0.109                 | 433                    | 0.55      | 0.82   | 0.66    |
| Low            | 1.064                 | 0.107                 | 520                    | 0.54      | 0.82   | 0.65    |
Table 4 (on next page)

Performance analyses of FunPred 3.0 with other protein function prediction methodologies
| Methods                                      | Precision | Recall | F-score |
|----------------------------------------------|-----------|--------|---------|
| FunPred 3.0                                  | 0.55      | 0.82   | 0.66    |
| FunPred-2 (Saha et al. 2017a)                | 0.51      | 0.90   | 0.65    |
| FPred_Apriori (Prasad et al., 2017)          | 0.64      | 0.66   | 0.65    |
| FunPred 1.1 (S. Saha et al., 2014)           | 0.61      | 0.50   | 0.55    |
| FunPred 1.2 (S. Saha et al., 2014)           | 0.63      | 0.56   | 0.59    |
| Deep_GO (Kulmanov et al. 2018)               | 0.48      | 0.49   | 0.48    |
| Chi square #1&2 (Hishigaki et al., 2001)     | 0.20      | 0.25   | 0.22    |
| Chi square #1 (Hishigaki et al., 2001)       | 0.25      | 0.27   | 0.26    |
| Neighborhood counting #1&2 (Schwikowski et al., 2000) | 0.28      | 0.41   | 0.33    |
| Neighborhood counting #1 (Schwikowski et al., 2000) | 0.26      | 0.45   | 0.33    |
| Fs-weight #1&2 (Chua et al., 2006)           | 0.36      | 0.43   | 0.39    |
| Fs-weight #1 (Chua et al., 2006)             | 0.33      | 0.42   | 0.37    |
| Nrc (Moosavi et al., 2013)                   | 0.37      | 0.43   | 0.40    |
| Zhang (S. Zhang et al., 2009)                | 0.20      | 0.19   | 0.19    |
| DCS (Peng et al., 2014)                      | 0.36      | 0.37   | 0.36    |
| DSCP (Peng et al., 2014)                     | 0.39      | 0.40   | 0.39    |
| PON (Liang et al., 2013)                     | 0.15      | 0.14   | 0.14    |
Table 5 (on next page)

Predicted samples of Unpredicted protein pair Interactions/functions ("Missing" protein-pair-Interactions/functions) in the MIPS dataset
| Interacting Protein Pairs | Predicted Interactions | Predicted Functions |
|--------------------------|------------------------|---------------------|
| **Protein#1** | **Protein#2** | **Interaction#1** | **Interaction#2** | **Function#1** | **Function#2** |
| YAL014c | YAL030w | two hybrid | coimmunoprecipitation | -- | -- |
| YAL014c | YMR197c | two hybrid | coimmunoprecipitation | -- | -- |
| YLR459w | YDR434w | Unable to Predict | -- | -- | -- |
| YDR167w | YBR081c | two hybrid | -- | -- | -- |
| YGL173c | YML085c | synthetic lethal | -- | -- | -- |
| YGL190c | YKL048c | synthetic lethal | two hybrid | Cell polarity | -- |
| YMR167w | YNL082w | coimmunoprecipitation | copurification | DNA repair | -- |
| YDR027c | YJR060w | affinity chromatography, affinity-tag GST | two hybrid | -- | -- |
| YGR082w | YNL131w | crosslinking | coimmunoprecipitation | -- | -- |
| YJR066w | YHR186c | synthetic lethal | -- | -- | Lipid metabolism |
| YDR363w-a | YER008c | synthetic lethal | -- | Vesicular transport | -- |
| YDR309c | YLR319c | synthetic lethal | Cell structure | -- | -- |
| YLR336c | YPL268w | Unable to Predict | -- | -- | -- |
| YKR099w | YDL106c | Unable to Predict | -- | -- | -- |
Table 6 (on next page)

Predicted samples of Unpredicted protein pair Interactions/functions (‘Unknown’ protein-pair-Interactions/functions) in the MIPS dataset
| Protein#1 | Protein#2 | Interaction#1       | Interaction#2 | Function#1          | Function#2          |
|----------|-----------|---------------------|---------------|---------------------|---------------------|
| YLR418c  | YIL040w   | two hybrid          | --            | Pol II Transcription| --                  |
| YOR326w  | YNL120c   | Mitosis             | --            | Cell polarity       | Cell cycle control  |
| YJR057w  | YDR438w   | Unable to Predict   | --            | --                  | --                  |
| YFL037w  | YMR299c   | Cell structure      | --            | RNA processing      | DNA repair          |
| YHR129c  | YGL124c   | Mitosis             | two hybrid    | --                  | --                  |
| YGR078c  | YAL011w   | synthetic lethal    | two hybrid    | --                  | --                  |
| YNL153c  | YDR149c   | two hybrid          | --            | Pol II transcription| --                  |
| YMR307w  | YMR317w   | two hybrid          | --            | Carbohydrate        | --                  |
| YLR039c  | YIL039w   | Vesicular transport | two hybrid    | --                  | --                  |
| YMR307w  | YHR004c   | two hybrid          | --            | Carbohydrate        | --                  |
| YDL003w  | YGL250w   | two hybrid          | --            | Energy generation   | --                  |
| YNL271c  | YGR228w   | Meiosis             | --            | Cell polarity       | Protein modification|
| YML094w  | YBR108w   | Unable to Predict   | --            | --                  | --                  |
| YEL003w  | YDR334w   | Unable to Predict   | --            | --                  | --                  |