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 [4,5]

[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 Plewczynski [4,5]

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.

59   Protein function prediction problem is characterized by several factors like the diversity
60 of members for functional groups, the hierarchical relationships among functional classes,
61 incomplete or missing information about proteins and their functions. Thus, it defines a complex
62 multi-label learning problem (Jiang & McQuay 2012; Valentini 2014; Zhang & Zhou 2014).
63 Hierarchical relationships among labels are described in MIPS Functional Catalogue and Gene
64 Ontology. Valentini (Valentini 2014) uses a binary classifier for each label according to True
65 Path Rule (TPR) and the funCat. Recent work of Guoxian and the co-authors (Yu et al. 2015),
66 explored the incomplete label problem in a hierarchical manner using function correlation.
67 Another approach for predicting protein function, as proposed by Piovesan et al. (Piovesan et al.
68 2015) includes the combination of the trio: PPIN information, protein domain and sequence. In
69 another work, Zhao et al. (Zhao et al. 2016) invokes dynamic weighted interaction network
70 instead of the static one. This dynamic network is enriched with PPIN, time course gene
71 expression data, protein's domain information and protein complex information which ultimately
72 predict function of a protein using majority ranking. While most of the predictive models
73 highlights on the most highly related similar proteins in the neighborhood of the test protein,
74 Reinders et al. (Reinders et al. 2018) focuses on the less similar proteins. It is shown by the
75 application of Label-Space Dimensionality Reduction (LSDR) techniques that though these
76 proteins are less similar but they are quite informative and plays an important role in protein
77 function prediction. Another iterative algorithm is implemented by Sun et al. (Sun et al. 2018)
78 for predicting protein functions. It is completely dependent on the identification of the functional
79 dependencies which are based on proteins and their interactions. Sequence similarity network is
80 another important aspect for protein function prediction, which is considered in the development
81 of the methodology, Effusion, as proposed in the work of Yunes & Babbitt (Yunes & Babbitt
82 2018). Other notable works in this field are Wang et al. (Wang et al. 2018) and Fa et al. (Fa et al.
83 2018)

84   All these methods discussed above have already taken protein function prediction to the
85 next higher level. Yet, the still uncovered details in the study and analysis support the need for
86 new computational methods exploiting the protein-protein interaction networks for biological
87 function identification. In our novel methodology FunPred 3.0, the functions of test proteins are
88 determined by analyzing the neighbourhood properties of their protein interaction network. At
89 the same time, certain selected physicochemical properties of amino acids are also used along
90 with it. This task is challenging because of several reasons; e.g., large number of functional
91 groups, different levels of the interconnection hierarchy, proteins with multiple functional
92 groups, and incomplete or missing labels. In this proposed methodology, MIPS dataset (Mewes
93 et al. 2002) is used. It contains protein pairs along with their corresponding functions. At the
94 initial phase, to estimate the effectiveness of FunPred 3.0, essential proteins are selected as test
95 proteins. The functions of these proteins are considered to be unknown for experimental purpose
96 though their functions are defined in the dataset. Then we have applied FunPred 3.0 to predict
97 the functions of the test proteins. Predicted functions are hence matched with the original ones to
98 compute precision, recall and F-Score. While executing FunPred 3.0, it has been observed that
99 870 PPIs out of 13528 PPIs i.e., ~6.4% of the overall MIPS dataset (Mewes et al. 2002) are
100 unpredicted i.e. either unknown or missing. FunPred 3.0 has been also applied to predict the
101 unannotated protein function and protein interaction (Mamoon et al. 2010), and also assigning
102 the functional annotations for 767 PPIs out of 870 PPIs, representing circa ~5.7% of the overall
103 MIPS dataset. Similar instances have been also observed in the case of subcellular location of
104 proteins where 1679 proteins out of 6721 unique number of proteins are still unpredicted i.e. the

105 subcellular localization of these proteins are still unknown. The predicted functional annotations
106 and subcellular localization of these unpredicted proteins and protein pairs respectively result in
107 relevant biological information, such as vital processes, diseases related mutations.

METHODOLOGY

109

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

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

133   PPIN formed of proteins and their corresponding interactions may contain essential/non-
134 essential proteins as well as reliable/unreliable edges. Proteins having maximum number of
135 interconnected neighbors are considered as essential proteins while proteins having less number
136 of interconnected neighbors are considered as non-essential ones. Presence of non-essential
137 neighbors in the PPIN might affect the unknown protein function prediction level accuracy. So
138 proper identification and elimination of non-essential proteins is needed to ensure the presence of
139 maximum number of essential proteins in the PPIN. In the proposed work, detection of essential
140 proteins is implemented by node weight. Node weight (Wang & Wu 2013) basically assigns a
141 weightage score to each node or protein based on its corresponding degree. High node weight
142 determines essential while low node weight detects non-essential protein. Thus non-essential
143 proteins are discarded from PPIN along their corresponding edges. Even after this initial phase of
144 PPIN refinement, there are still some unreliable edges present in the network. Since protein
145 clusters are formed, so it is obvious that two nodes with an edge between them belong to the
146 same cluster if they have high similarity. Edge between two nodes of high similarity is

147 considered as reliable edge while that of low similarity is denoted as unreliable edge. In the
148 proposed work, detection of reliable edges is executed by edge weight. Edge weight (Wang &
149 Wu 2013) also assigns a weightage score to each edge connected by two proteins in the
150 terminals. Assignment of edge weight to an edge depends on the number of common neighbors
151 between the two terminal proteins of the corresponding edge. More number of common
152 neighbors signifies high similarity which in turn detects reliable edges. On the other hand,
153 unreliable edges have low similarity since they have less number of common neighbors. Thus,
154 unreliable edges are identified and pruned from the PPIN. Filtered out PPIN after these
155 refinements, contains only essential proteins and reliable edges, which ultimately helps in
156 enhancing prediction accuracy level.

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

163 1. All the unique proteins are first clustered into M mutually exclusive clusters based on
164   their node weight and edge weight in the overall PPIN. Node and edge weight have been
165   used to ensure all the most essential nodes with higher reliability are present in the cluster
166   and get selected as test proteins.
167
168 2. Functional annotations are then derived from the multi-level neighbourhood of an
169   unknown protein within each cluster.

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

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

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

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

186 The entire methodology has been described in Algorithm 1 as well as pictorially
187 highlighted in Figure 1 and Figure 2. In Figure 1, sample Table 1 (node weight table) is formed

188 from the initial PPIN of yeast. Hence, node weight threshold is calculated using equation 1 at
189 three levels: high, medium and low. These thresholds are applied on the initial PPIN to filter out
190 three sub-networks i.e. sub-network1, sub-network2, sub-network3 respectively at high, medium
191 and low node weight thresholds. Respective edge weight tables i.e. sample Table 2 (in Figure 1),
192 sample Table 3 (in Figure 1), sample Table 4 (in Figure 1) are formed from sub-network1, sub-
193 network2, sub-network3, upon which high, medium and low edge weight thresholds (obtained
194 using equation 1) are applied to form pruned sub-network4, sub-network5 and sub-network3.

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

207  All the proteins belonging to M mutually exclusive protein clusters (under three levels of
208 thresholds: high, medium and low separately), obtained from FunPred 3.0_Clust are considered
209 as essential proteins and hence they are included in the test set of our proposed methodology. In
210 the second step, FunPred 3.0_Pred predicts labels these selected unlabelled (or test) proteins
211 using neighbourhood properties and physicochemical properties of amino acids (see Algorithm 2
212 and Figure 3). In Figure 3, for each test protein (say P_1 belonging to refined sub-network4 in
213 Figure 1) its *level* 1 neighborhood graph is formed including those proteins which are present in
214 its corresponding node weight table: sample Table 1 under high threshold in Figure 1. Then
215 protein clusters are formed at *level* 1 (considering each level 1 protein as the seed of the cluster)
216 in a similar way as formed in FunPred 3.0_Clust. It should be noted here that the number of
217 clusters formed here is equivalent to the number of proteins in *level* 1. Distance between the
218 mean of the physico-chemical features of each protein cluster as well as test protein is computed
219 and the test belongs to the cluster having the least distance. All the functions of the selected
220 cluster are allocated to the test protein.

221  The relevant $level-1$ neighbours of the test proteins are chosen to form their individual
222 neighbourhood graph. In finding the $level-1$ neighbours or forming their individual
223 neighbourhood graph, relevance is measured in terms of edge weight properties. Next, PCP
224 score is computed for the neighbourhood graph of each test protein. Six different high-ranked
225 physico-chemical features: aliphatic index (Singh et al. 2008), gravy (Kyte & Doolittle 1982),
226 aromacity (Lobry & Gautier 1994), number of negatively charged residues (Singh et al. 2008),
227 number of positively charged residues (Singh et al. 2008), isoelectric point (Bjellqvist et al.
228 1994) are used to reckon this physico-chemical property (PCP) based score. These high-ranked
229 features are selected from ten divergent physico-chemical features (see supplementary) by the
230 enactment of four distinct classifiers: XGBoost classifier (Chen & Guestrin 2016; Pedregosa et
231 al. 2011), Random Forest classifier (Breiman 2001; Pedregosa et al. 2011), Extra Tree classifier

232 (Geurts et al. 2006; Pedregosa et al. 2011) and Recursive feature elimination classifier
233 (Pedregosa et al. 2011). High-ranked five among ten features have been picked at first by each
234 classifier. Then from these picked features, frequency of maximum occurrences for each
235 individual feature has been noted from which endmost six features get selected (see Table 1).
236 Finally, each test protein is assigned to a functional group of the neighbourhood graph, based on
237 the nearest neighbourhood approach on the basis of mean PCP score. FunPred 3.0_Clust (see
238 Algorithm 1) and FunPred 3.0_Pred (see Algorithm 2) describes the methodology of unknown
239 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_{high}', G_{medium}' and G_{low}' from G

    G_{high}', G_{medium}' and G_{low}' consisting of only essential proteins (high node weight) are formed

    //computation of edge weight of G'

    for all edges in G_{high}', G_{medium}' and G_{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_{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_{medium}' and G_{low}' respectively.

    //Formation of refined sub-networks $G_{high,high}''$, $G_{medium,medium}''$ and $G_{low,low}''$ from G_{high}', G_{medium}' and G_{low}' at high node and edge weight threshold, medium node and edge weight threshold, low node and edge weight threshold respectively.

    form $G_{high,high}''$, $G_{medium,medium}''$ and $G_{low,low}''$ consisting of only reliable edges (high edge weight)

    //Formation of clusters at three levels of thresholds

    for all proteins in $G_{high,high}''$

        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^{high}

    i.e. $C_M^{high} = \{P\}$ where 1=<M<=W (W is the total no. of nodes in node weight   table)

    neighbors of P are added to C_M^{high} provided its inclusion does not cause edge weight to fall below high edge weight threshold value i.e. $C_M^{high} = \{P\} \cup N_{P_1}$

    update the node edge table by eliminating all the proteins present in C_M^{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 $G_{medium,medium}''$ and $G_{low,low}''$.

End

240

Algorithm 2. FunPred 3.0_Pred:
(Protein function prediction of test proteins)

Input:	Set of un-annotated proteins in C_M^{high}, C_M^{medium}, C_M^{low} 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_M^{high}

        for each $level-1$ neighbor N of P present in $G_{high}^{'}$

          add N as the seed of the cluster K_i i.e. K_i={N}.

          //where 1=<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_M^{high}

        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_M^{high}

      for each protein P in C_M^{high}

        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_M^{medium} and C_M^{low}

End.

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

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

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

283   Subcellular localization dataset of yeast contains 1679 number of unpredicted localization
284   proteins out of which our method predicts the localization of 638 proteins successfully.

285 Localization information for the remaining 1041 proteins cannot be predicted because of the
286 absence of PPIN interaction in MIPS dataset as discussed earlier. This extra added layer of
287 biological information about subcellular localization of proteins along with the protein function
288 prediction boost up our methodology FunPred 3.0 to the next higher level.

289 **RESULTS**

290

291 Initially, PPIN of yeast consists of 4554 unique proteins in 13528 protein-protein interactions
292 (PPINs) after the elimination of the self-replicating and the self-interacting protein pairs. After
293 the network refinement through the execution of node and edge weight threshold, non-essential
294 proteins along with unreliable edges get eliminated and the initial PPIN gets reduced to almost
295 3174 unique proteins and 6936 PPINs (approx.) considering three levels of thresholds from
296 which FunPred 3.0_Clust form protein clusters to generate test set of proteins.

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

309 $$\text{P} = \frac{TP}{TP+FP} \quad \text{R} = \frac{TP}{TP+FN} \quad \text{F} = \frac{2*(P*R)}{P+R}$$

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

311 The performance of FunPred 3.0 has been analyzed under different levels of thresholds of
312 node and edge weights as highlighted in Table 3. It should be noted here that under high and medium
313 thresholds, same Precision, Recall and F-score have been retrieved since number of selected test
314 proteins are equivalent in both the cases. All over result analysis as depicted in Table 3 shows that
315 there are not much significant changes in the result as such varying the thresholds except a slight fall
316 in precision and F-score under low threshold as compared to the others. The fall is due to the
317 relaxation in the node and edge weight thresholds resulting in incorporation of less essential proteins
318 in the test set. So, the high threshold ensures inclusion of the most essential proteins. So the
319 Precision, Recall and F-score of FunPred 3.0 are reckoned as 0.55, 0.82 and 0.66 respectively under
320 high threshold. High Recall and low Precision emerges out as a major characteristic of FunPred 3.0
321 when compared to the other methodologies except FunPred-2. It highlights the fact that most of
322 the admissible results are successfully generated by FunPred 3.0. Table 4 shows a detailed
323 performance comparison of other methodologies along with our proposed systems (like FunPred
324 1.1, FunPred 1.2 (Saha et al. 2014)), neighbourhood counting method (Schwikowski et al. 2000), the

325   chi-square method (Hishigaki et al. 2001), a recent version of the neighbor relativity coefficient
326   (NRC) (Moosavi et al. 2013), FPred_Apriori (Prasad et al. 2017), Zhang methodology (Zhang et al.
327   2009), Domain Combination Similarity (DCS) (Peng et al. 2014), Domain Combination Similarity
328   in context of protein complexes (DSCP) (Peng et al. 2014), Protein Overlap Network (PON) (Liang
329   et al. 2013), Deep_GO (Kulmanov et al. 2018) and the FS-weight based method (Chua et al. 2006)).
330   All these data are collected from their respective works which are executed on the same organism i.e.
331   yeast. The results of Deep_GO (Kulmanov et al. 2018) are computed manually for yeast dataset, the
332   code of which is available at https://github.com/SovanSaha/FunPred-3.0.git. From Table 4, it can
333   be also highlighted that our method, FunPred 3.0, yields relatively higher F-Score values than the
334   others including its earlier version FunPred-2.

335   Table 4 also discloses the fact that NRC method has overpowered the rest except FunPred
336   1.1 (Saha et al. 2014), FunPred 1.2 (Saha et al. 2014), FPred_Apriori (Prasad et al. 2017) and
337   FunPred 3.0. The reason behind this is observed as follows: Both version of FunPred 1 has
338   incorporated two levels (i.e. $level-1$ and $level-2$) of PPIN as well as lot of essential
339   neighborhood properties like neighborhood ratio, protein path connectivity and relative functional
340   similarity (includes both ancestor and descendant information of a specific protein) have been
341   utilized to assess the reliability of each node (protein) along with its associated edges (protein
342   interaction) during the unannotated protein function prediction. FPred_Apriori (Prasad et al. 2017)
343   executes both closeness centrality and edge clustering coefficient to make its predictive approach
344   more effective than the others. Last but not least, FunPred 3.0 combines physico-chemical property
345   of each protein along with neighborhood analysis (like node weight, edge weight etc.) for predicting
346   protein function which ultimately promotes it to the next higher level in the terms of performance
347   analysis when compared to the others.

348   Though Neighborhood counting method is simple in nature yet the performance measure of
349   it has descended considerably in comparison to NRC (Moosavi et al. 2013), FS-weight #1 (directly
350   connected proteins) and FS-weight #1 and #2 (directly and indirectly connected proteins) despite of
351   its simplicity (Chua et al. 2006). This is because no differentiation has been observed between the
352   direct and indirect neighborhood connection. Beside most of the methods included in Table 4 like
353   NRC (Moosavi et al. 2013), Chi square #1&2 (Hishigaki et al. 2001), Chi square #1 (Hishigaki et al.
354   2001), Neighborhood counting #1&2 (Schwikowski et al. 2000), Neighborhood counting #1
355   (Schwikowski et al. 2000) etc. are not utilized for the refinement of the PIN by pruning unreliable
356   proteins or edges which in turn increases false positives in their prediction accuracy level. In
357   FPred_Apriori (Prasad et al. 2017), a bottom-up predictor of existing Apriori algorithm has been
358   utilized for protein function prediction by exploiting two most important neighborhood properties:
359   closeness centrality and edge clustering coefficient of protein interaction network. Though the
360   method is unique in the fact that the functions of the leaf nodes in the interaction network have been
361   back propagated and thus labeled up to the root node (test protein) but yet it fails to generate high
362   Recall and F-score than FunPred 3.0. But it returns substantially high precision values than the
363   others as well as all our methods. DCS (Peng et al. 2014), DSCP (Peng et al. 2014), PON (Liang et
364   al. 2013), Deep_GO (Kulmanov et al. 2018) and Zhang methodology (Zhang et al. 2009) are well
365   developed methods for protein function prediction incorporating domain specific as well as
366   neighborhood based properties but they fail to compete with all our methodologies due to the lack of
367   important feature selection methodologies of physicochemical properties and proper assessment of
368   nodes and edges involved in test set protein function prediction through node and edge weights.

369   During experimental evaluation, the validation set is prepared with 4554 labeled
370 *Saccharomyces cerevisiae* proteins, collected from the MIPS dataset. Using FunPred 3.0_Clust,
371 we identify M mutually exclusive protein clusters (Wang & Wu 2013). Experimental variations
372 with k = 1,2,3, are included in Table 2. Using an optimal choice of k = 3, we identify 433 test
373 targets for the validation set. Now, the functional labels of these test proteins are assigned using
374 FunPred 3.0_Pred. The Precision, Recall, F-scores of our method over the test targets of the
375 validation set is obtained as 0.55, 0.82 and 0.66 respectively.

DISCUSSION

377

378 Our results (characterized by the Precision, Recall and F-Score) and comparison with the other
379 protein functional group prediction models show the superiority of our approach. The FunPred
380 3.0 software has better performance than any existing function prediction *in silico* method. The
381 network structure may be pruned based on the edge weight and along with it use of physico-
382 chemical properties lead to improved and faster functional prediction in complex and diverse
383 protein–protein interaction networks. We would like to estimate the effectiveness of our in silico
384 method for other organisms, such as in human protein–protein interactions with even more
385 complex network architectures.

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

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

CONCLUSION

406

407   FunPred 3.0 thus proved to be an improved and advanced version of our previous

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

427 **REFERENCES**
428

429 Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M,
430   Martin MJ, Natale DA, O'Donovan C, Redaschi N, and Yeh L-SL. 2004. UniProt: the Universal Protein
431   knowledgebase. *Nucleic Acids Research* 32(Database issue):D115-D119. 10.1093/nar/gkh131.
432

433 Bjellqvist B, Basse B, Olsen E, and Celis JE. 1994. Reference points for comparisons of two-dimensional maps of
434   proteins from different human cell types defined in a pH scale where isoelectric points correlate with
435   polypeptide compositions. *ELECTROPHORESIS* 15(1):529-539. 10.1002/elps.1150150171.
436

437 Breiman L. 2001. Random Forests. *Machine Learning* 45(1):5-32. 10.1023/a:1010933404324.
438

439 Chakicherla A, Ligon Dang M, Rodriguez V, Hansen J, and Zemla A. 2018. *Function prediction of an interacting*
440   *protein pair using domain fusion analysis: SpaR and SpaK*.
441

442 Chatterjee P, Basu S, Kundu M, Nasipuri M, and Plewczynski D. 2011a. PPI_SVM: Prediction of protein-protein
443   interactions using machine learning, domain-domain affinities and frequency tables. *Cellular & Molecular*
444   *Biology Letters* 16(2):264-278. 10.2478/s11658-011-0008-x.
445

446 Chatterjee P, Basu S, Kundu M, Nasipuri M, and Plewczynski D. 2011b. PSP_MCSVM: brainstorming consensus
447   prediction of protein secondary structures using two-stage multiclass support vector machines. *Journal of*
448   *Molecular Modeling* 17(9):2191. 10.1007/s00894-011-1102-8.
449

450 Chatterjee T, Chatterjee P, Basu S, Kundu M, and Nasipuri M. 2012. Protein function by minimum distance
451   classifier from protein interaction network. In: 2012 International Conference on Communications, Devices
452   and Intelligent Systems (CODIS).588-591.
453

454 Chen J, Hsu W, Lee ML, and Ng SK. 2007. Labeling network motifs in protein interactomes for protein function
455   prediction. In: 2007 IEEE 23rd International Conference on Data Engineering.546-555.

456
457 Chen T, and Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM
458   SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California,
459   USA: ACM.785-794.
460
461 Chua HN, Sung W-K, and Wong L. 2006. Exploiting indirect neighbours and topological weight to predict protein
462   function   from   protein–protein   interactions.   *Bioinformatics*   22(13):1623-1630.
463   10.1093/bioinformatics/btl145.
464
465 Fa R, Cozzetto D, Wan C, and Jones DT. 2018. Predicting human protein function with multi-task deep neural
466   networks. *PLOS ONE* 13(6):e0198216. 10.1371/journal.pone.0198216.
467
468 Geurts P, Ernst D, and Wehenkel L. 2006. Extremely randomized trees. *Machine Learning* 63(1):3-42.
469   10.1007/s10994-006-6226-1.
470
471 Hishigaki H, Nakai K, Ono T, Tanigami A, and Takagi T. 2001. Assessment of prediction accuracy of protein
472   function from protein–protein interaction data. *Yeast* 18(6):523-531. 10.1002/yea.706.
473
474 Jiang JQ, and McQuay LJ. 2012. Predicting Protein Function by Multi-Label Correlated Semi-Supervised Learning.
475   *IEEE/ACM   Transactions   on   Computational   Biology   and   Bioinformatics*   9(4):1059-1069.
476   10.1109/TCBB.2011.156.
477
478 Kulmanov M, Khan MA, Hoehndorf R, and Wren J. 2018. DeepGO: predicting protein functions from sequence and
479   interactions using a deep ontology-aware classifier. *Bioinformatics (Oxford, England)* 34(4):660-668.
480   10.1093/bioinformatics/btx624.
481
482 Kyte J, and Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of
483   Molecular Biology* 157(1):105-132. https://doi.org/10.1016/0022-2836(82)90515-0.
484
485 Lee D, Redfern O, and Orengo C. 2007. Predicting protein function from sequence and structure. *Nature Reviews
486   Molecular Cell Biology* 8(995. 10.1038/nrm2281
487
488 Liang S, Zheng D, Standley DM, Guo H, and Zhang C. 2013. A novel function prediction approach using protein
489   overlap networks. *BMC Systems Biology* 7(1):61. 10.1186/1752-0509-7-61.
490
491 Lichtarge O, Bourne HR, and Cohen FE. 1996. An Evolutionary Trace Method Defines Binding Surfaces Common
492   to Protein Families. *Journal of Molecular Biology* 257(2):342-358. https://doi.org/10.1006/jmbi.1996.0167.
493
494 Lobry JR, and Gautier C. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid
495   usage in 999 Escherichia coli chromosome-encoded genes. *Nucleic Acids Research* 22(15):3174-3180.
496
497 Mamoon R, Sumathy R, and Gajendra PSR. 2010. A Simple Approach for Predicting Protein-Protein Interactions.
498   *Current Protein & Peptide Science* 11(7):589-600. http://dx.doi.org/10.2174/138920310794109120.
499
500 Mewes HW, Frishman D, Güldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Münsterkötter M,
501   Rudd S, and Weil B. 2002. MIPS: a database for genomes and protein sequences. *Nucleic Acids Research*
502   30(1):31-34.
503
504 Mills CL, Beuning PJ, and Ondrechen MJ. 2015. Biochemical functional predictions for protein structures of
505   unknown or uncertain function. *Computational and Structural Biotechnology Journal* 13(182-191.
506   https://doi.org/10.1016/j.csbj.2015.02.003.
507
508 Moosavi S, Rahgozar M, and Rahimi A. 2013. Protein function prediction using neighbor relativity in protein–
509   protein   interaction   network.   *Computational   Biology   and   Chemistry*   43(11-16.
510   https://doi.org/10.1016/j.compbiolchem.2012.12.003.

511   Ng PC, and Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids*
512   *Research* 31(13):3812-3814.
513
514   Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R,
515   Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, and Duchesnay E. 2011. Scikit-
516   learn: Machine Learning in Python. *J Mach Learn Res* 12(2825-2830.
517
518   Peng W, Wang J, Cai J, Chen L, Li M, and Wu F-X. 2014. Improving protein function prediction using domain and
519   protein complexes in PPI networks. *BMC Systems Biology* 8(1):35. 10.1186/1752-0509-8-35.
520
521   Piovesan D, Giollo M, Leonardi E, Ferrari C, and Tosatto SCE. 2015. INGA: protein function prediction combining
522   interaction networks, domain assignments and sequence similarity. *Nucleic Acids Research* 43(Web Server
523   issue):W134-W140. 10.1093/nar/gkv523.
524
525   Prasad A, Saha S, Chatterjee P, Basu S, and Nasipuri M. 2017. Protein Function Prediction from Protein Interaction
526   Network Using Bottom-up L2L Apriori Algorithm. In: Computational Intelligence, Communications, and
527   Business Analytics. Singapore: Springer Singapore.3-16.
528
529   Reinders MJT, van Ham RCHJ, and Makrodimitris S. 2018. Improving protein function prediction using protein
530   sequence and GO-term similarities. 10.1093/bioinformatics/bty751.
531
532   Saha S, Chatterjee P, Basu S, Kundu M, and Nasipuri M. 2012. Improving prediction of protein function from
533   protein interaction network using intelligent neighborhood approach. In: 2012 International Conference on
534   Communications, Devices and Intelligent Systems (CODIS).584-587.
535
536   Saha S, Chatterjee P, Basu S, Kundu M, and Nasipuri M. 2014. FunPred-1: Protein function prediction from a
537   protein interaction network using neighborhood analysis. *Cellular & Molecular Biology Letters* 19(4):675-
538   691. 10.2478/s11658-014-0221-5.
539
540   Saha S, Chatterjee P, Basu S, and Nasipuri M. 2017a. Functional Group Prediction of Un-annotated Protein by
541   Exploiting Its Neighborhood Analysis in Saccharomyces Cerevisiae Protein Interaction Network. In: Chaki
542   R, Saeed K, Cortesi A, and Chaki N, eds. *Advanced Computing and Systems for Security: Volume Four*.
543   Singapore: Springer Singapore, 165-177.
544
545   Saha S, Sengupta K, Chatterjee P, Basu S, and Nasipuri M. 2017b. Analysis of protein targets in pathogen–host
546   interaction in infectious diseases: a case study on Plasmodium falciparum and Homo sapiens interaction
547   network. *Briefings in Functional Genomics*:elx024-elx024. 10.1093/bfgp/elx024.
548
549   Schwikowski B, Uetz P, and Fields S. 2000. A network of protein–protein interactions in yeast. *Nature*
550   *Biotechnology* 18(1257. 10.1038/82360
551   Shatsky M, Allen S, Gold BL, Liu NL, Juba TR, Reveco SA, Elias DA, Prathapam R, He J, Yang W, Szakal ED,
552   Liu H, Singer ME, Geller JT, Lam BR, Saini A, Trotter VV, Hall SC, Fisher SJ, Brenner SE, Chhabra SR,
553   Hazen TC, Wall JD, Witkowska HE, Biggin MD, Chandonia J-M, and Butland G. 2016. Bacterial
554   Interactomes: Interacting Protein Partners Share Similar Function and Are Validated in Independent Assays
555   More Frequently Than Previously Reported. *Molecular & Cellular Proteomics : MCP* 15(5):1539-1555.
556   10.1074/mcp.M115.054692.
557
558   Singh M, Wadhwa PK, and Kaur S. 2008. Predicting Protein Function using Decision Tree. *World Academy of*
559   *Science, Engineering and Technology* 2(3):300-303.
560
561   Sriwastava BK, Basu S, and Maulik U. 2015. Predicting Protein-Protein Interaction Sites with a Novel Membership
562   Based Fuzzy SVM Classifier. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*
563   12(6):1394-1404. 10.1109/TCBB.2015.2401018.
564

565 Sun P, Tan X, Guo S, Zhang J, Sun B, Du N, Wang H, and Sun H. 2018. Protein Function Prediction Using
566   Function Associations in Protein–Protein Interaction Network. *IEEE Access* 6(30892-30902.
567   10.1109/ACCESS.2018.2806478.

569 Valentini G. 2014. Hierarchical Ensemble Methods for Protein Function Prediction. *ISRN Bioinformatics* 2014(34.
570   10.1155/2014/901419.

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

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

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

580 Xiong W, Liu H, Guan J, and Zhou S. 2013. Protein function prediction by collective classification with explicit and
581   implicit edges in protein-protein interaction networks. *BMC Bioinformatics* 14(Suppl 12):S4-S4.
582   10.1186/1471-2105-14-S12-S4.

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

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

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

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

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

599 Zhao B, Wang J, Li M, Li X, Li Y, Wu FX, and Pan Y. 2016. A New Method for Predicting Protein Functions From
600   Dynamic Weighted Interactome Networks. *IEEE Transactions on NanoBioscience* 15(2):131-139.
601   10.1109/TNB.2016.2536161.

602

Figure 1

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

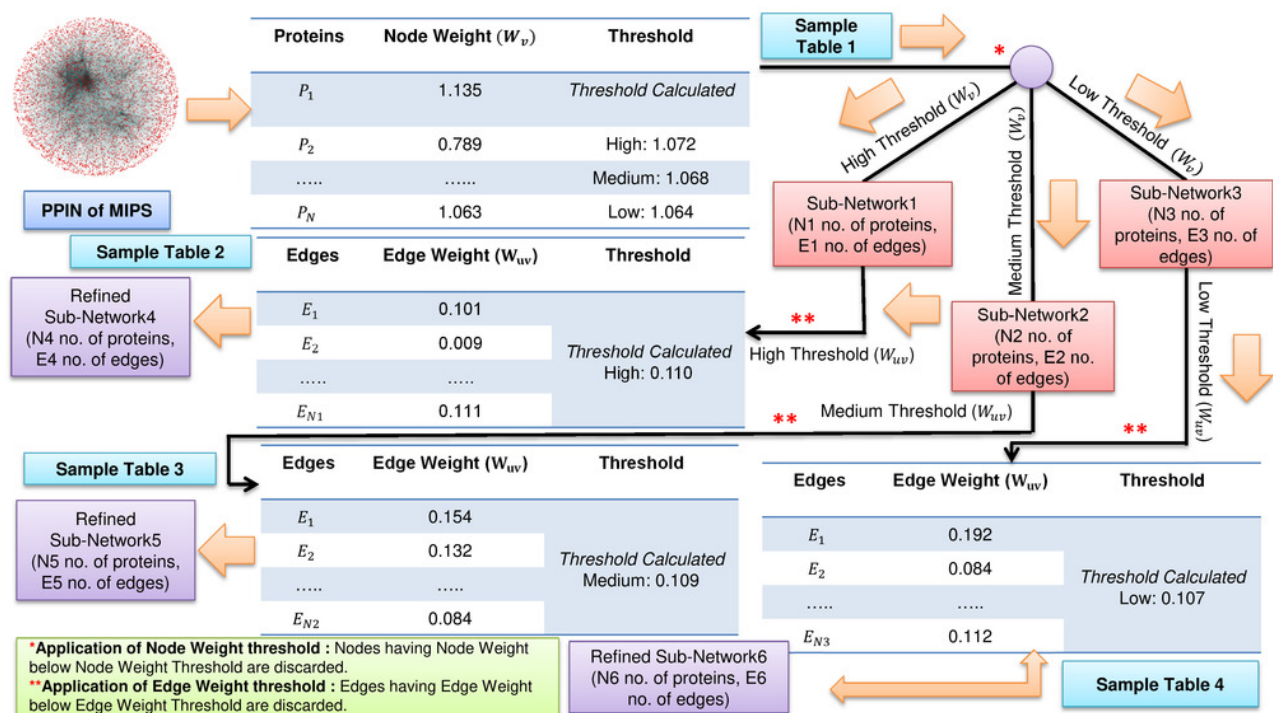



Figure 2

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_1 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_1.	
5	Delete all proteins in the cluster of P_1 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_Pred

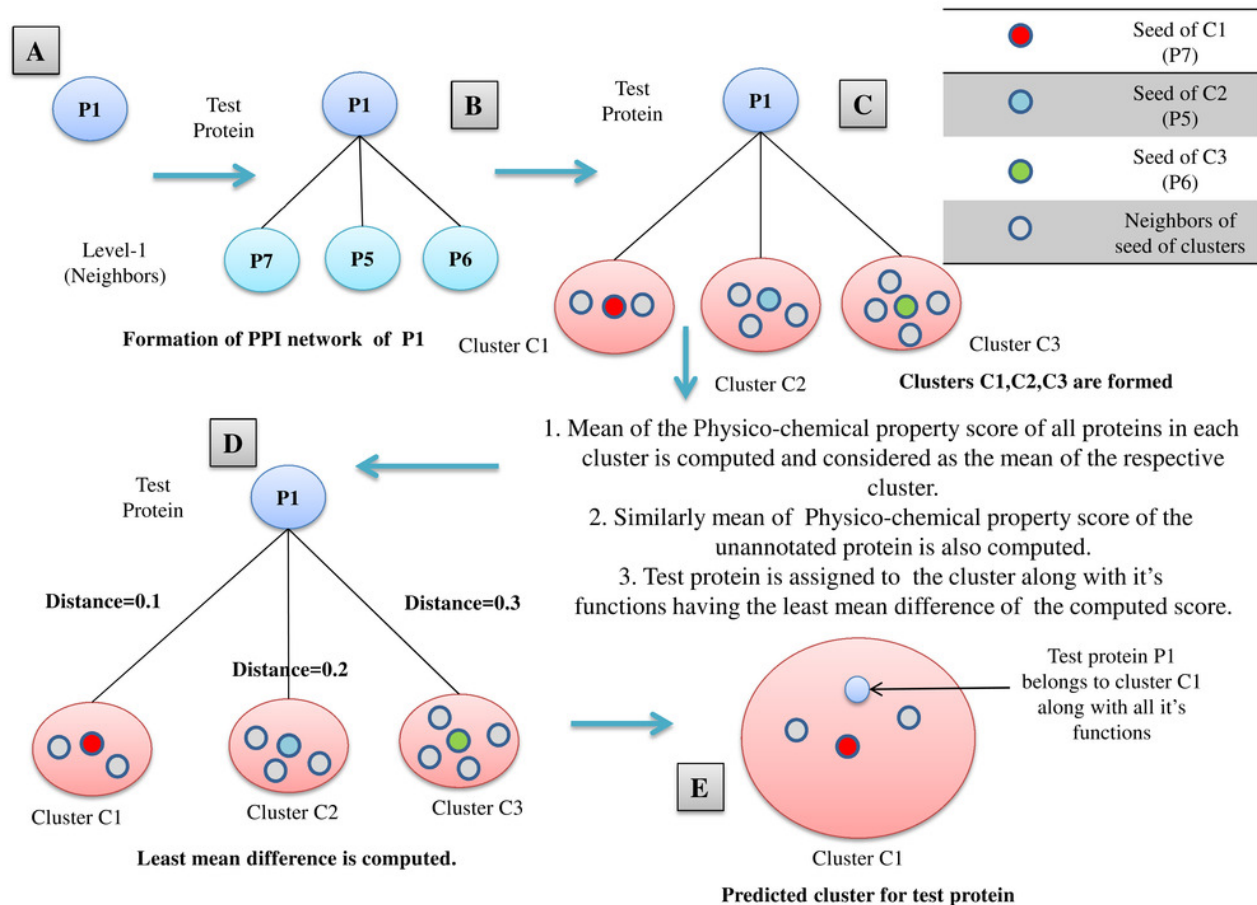



Figure 4

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

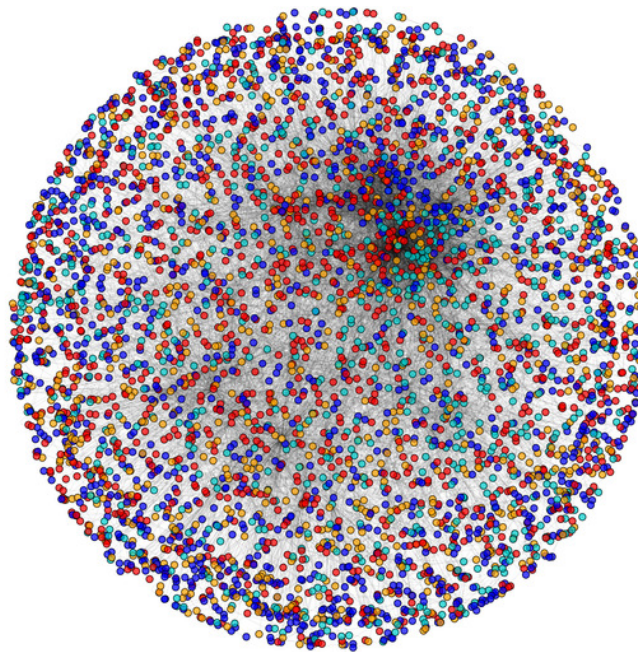



Figure 5

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

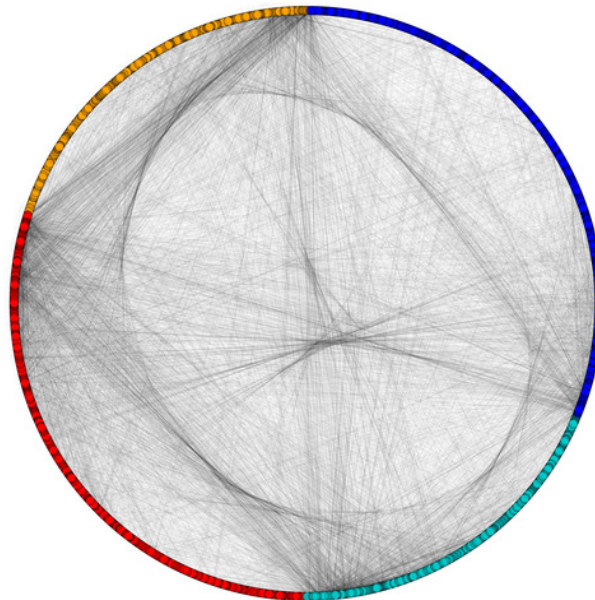



Figure 6

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

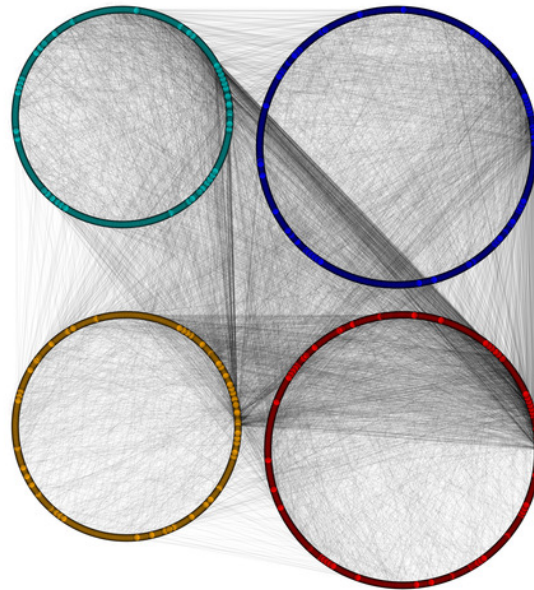



Figure 7

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

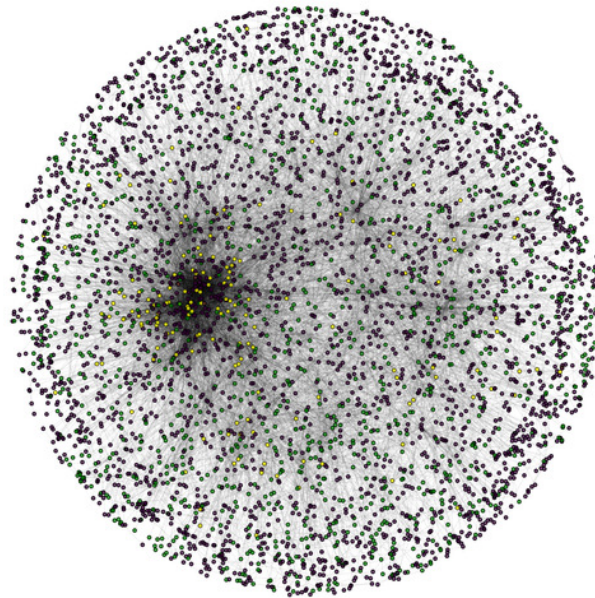



Figure 8

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

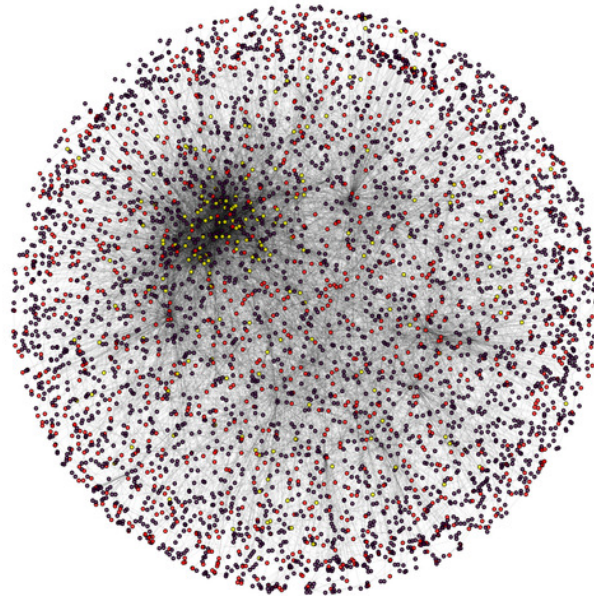



Figure 9

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

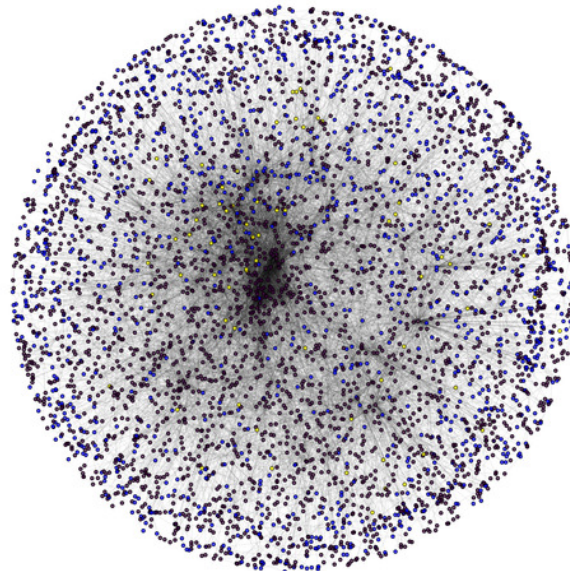



Figure 10

PPI network of Yeast (*Saccharomyces cerevisiae*)

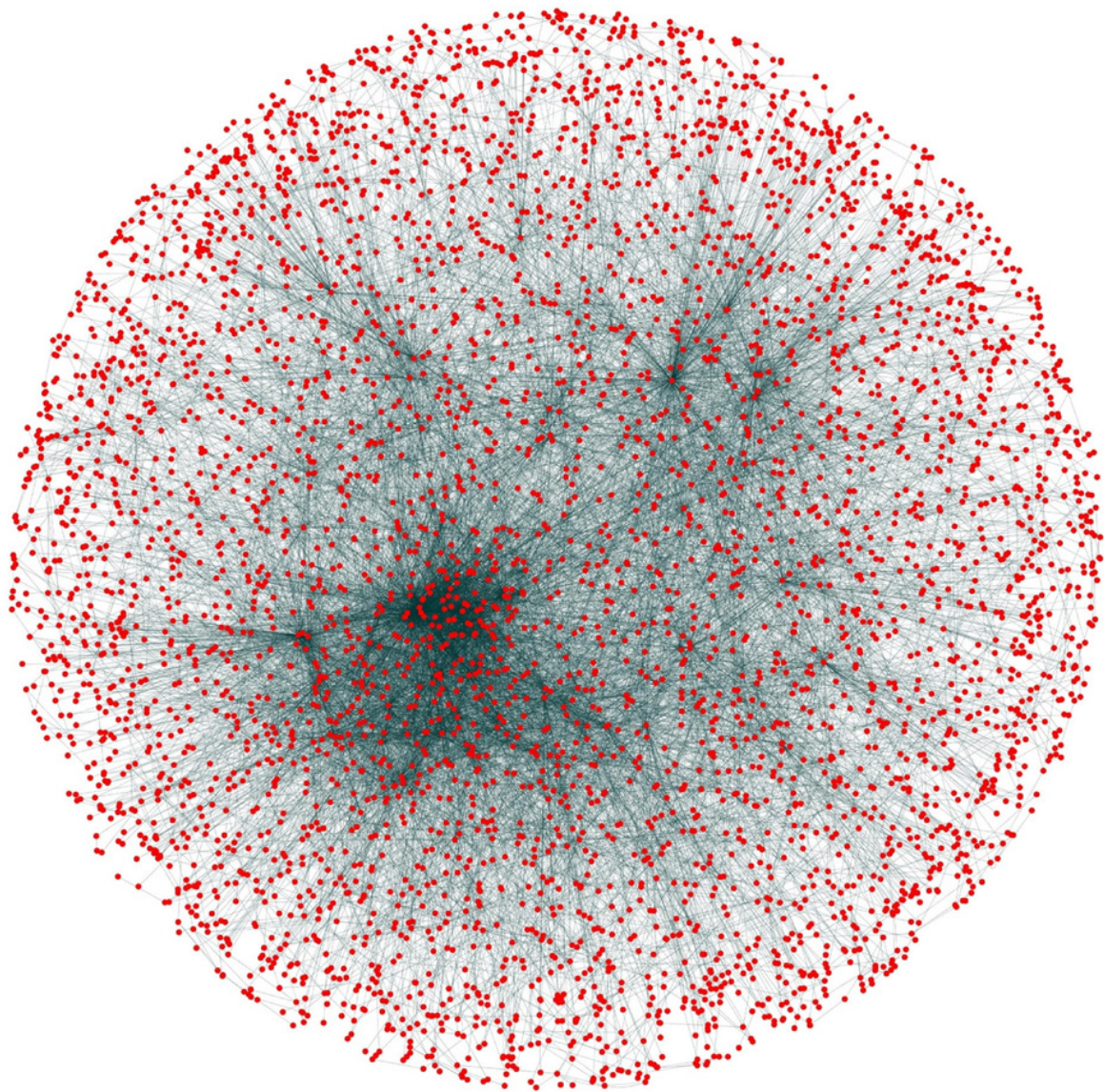
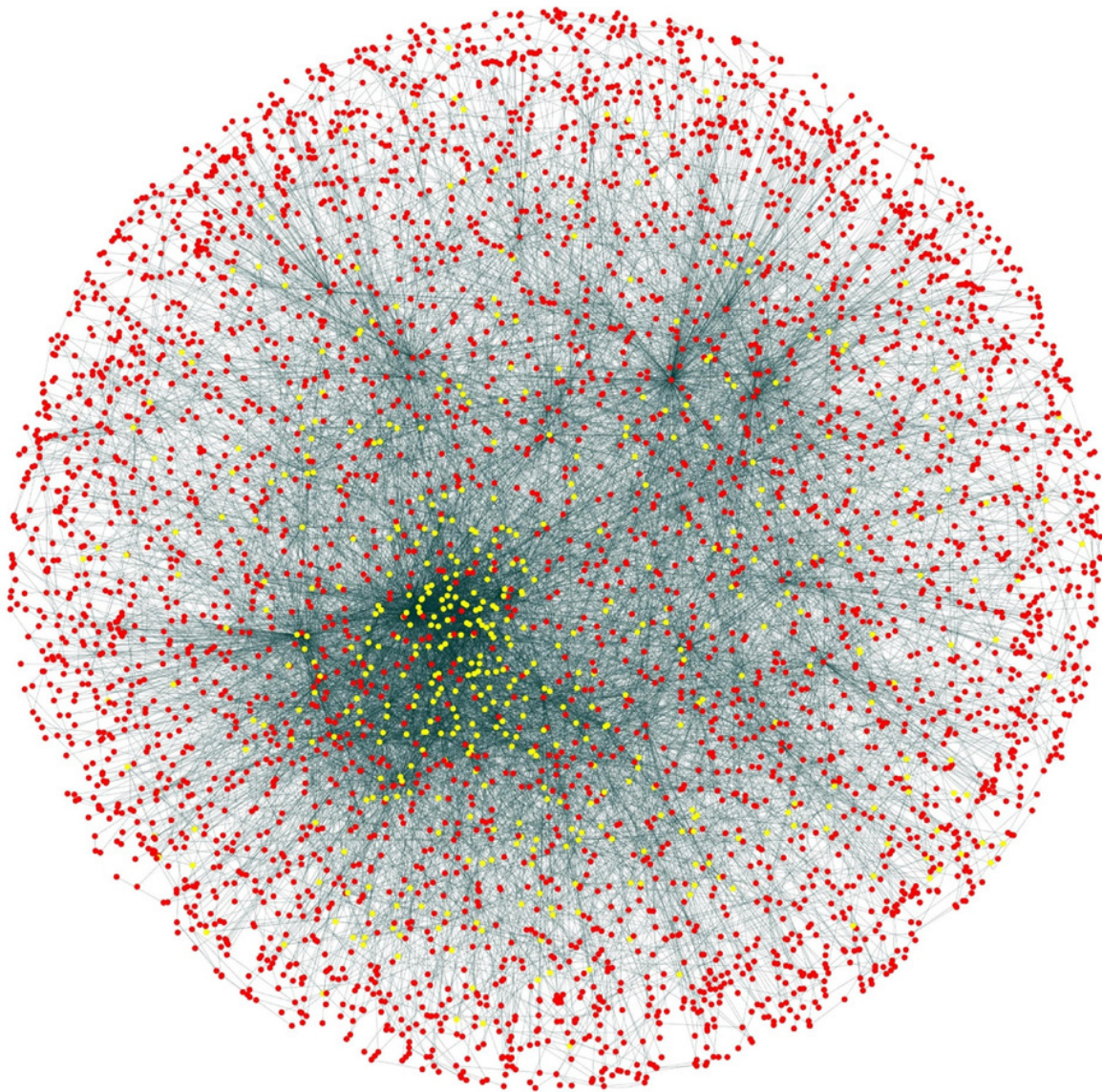



Figure 11

PPI network of annotated (red circle) and test/unannotated proteins (yellow circle) of Yeast network (*Saccharomyces cerevisiae*)

Top-ranked selected physicochemical features (marked in blue) using 4 classifiers based on the maximum number of hits

Physicochemical properties	Classifiers Used (Returns #5 top-ranked physicochemical properties/features)				
	XGBoost	Random Tree	Extra Tree	Recursive Feature Elimination	#Hits
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

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

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

2

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

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	--	--	--

Predicted samples of Unpredicted protein pair Interactions/functions ("Unknown" 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**
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 metabolism	--
YLR039c	YIL039w	Vesicular transport	two hybrid	--	--
YMR307w	YHR004c	two hybrid	--	Carbohydrate metabolism	--
YDL003w	YGL250w	two hybrid	--	Energy generation	--
YNL271c	YGR228w	Meiosis	--	Cell polarity	Protein modification
YML094w	YBR108w	Unable to Predict	--	--	--
YEL003w	YDR334w	Unable to Predict	--	--	--