Manchester Metropolitan University

e-space

Manchester Metropolitan University's Research Repository

Please cite the Published Version

Ur Rehman, Hafeez, Habib, Usman, Ijaz, Umer, Islam, Naveed, Khan, Atta Ur Rehman and Nawaz, Raheel (2019) Identification of Yeast's Interactome Using Neural Networks. IEEE Access, 7. pp. 179634-179645.

DOI: https://doi.org/10.1109/access.2019.2959401

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Version: Published Version

Downloaded from: https://e-space.mmu.ac.uk/624662/

Usage rights: CC BY Creative Commons: Attribution 4.0

Additional Information: This is an Open Access article published in IEEE Access, published by IEEE, copyright The Author(s).

Enquiries:
If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines)

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 31, 2019, accepted December 2, 2019, date of publication December 13, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2959401

Identification of Yeast's Interactome Using Neural Networks

HAFEEZ UR REHMAN[1], USMAN HABIB[1], UMER IJAZ[2], NAVEED ISLAM[3], ATTA UR REHMAN KHAN[4], (Senior Member, IEEE), AND RAHEEL NAWAZ[5]

[1]Department of Computer Science, FAST National University of Computer and Emerging Sciences—Peshawar, Peshawar 24720, Pakistan
[2]Department of Electrical Engineering, Government College University, Faisalabad 38000, Pakistan
[3]Department of Computer Science, Islamia College University, Peshawar 25120, Pakistan
[4]Faculty of Computing and Information Technology, Sohar University, Sohar 311, Oman
[5]Business School, Manchester Metropolitan University (MMU), Manchester M15 6BH, U.K.

Corresponding author: Hafeez Ur Rehman (hafeez.urrehman@nu.edu.pk)

ABSTRACT An important aspect for designing precise medical therapies is to have an accurate knowledge of protein protein interactions involved in the process. Next generation sequencing technologies for discovering novel genes are fuelling an information explosion that allows researchers to study these molecules in previously unimagined ways. A profound understanding of these biological components promises great leaps for the field of medical sciences, i.e., by designing personalized/preventive medicines, or by curing the life threatening diseases including many types of cancers etc. However, experimental techniques are slow and noisy; and usually report false interactions. These limitations prompted a surge of interest in computational techniques to infer the true interactions. In this paper, we propose and evaluate a Neural Network based approach for deciphering the interactions of *Saccharomyces cerevisiae* species proteins. The novelty of this approach lies in integrating the evidences (broadly classified as structural and non-structural) in a hybrid fashion. The structure-based evidences include, geometrical features extracted from individual homolog templates, e.g., interacting residues, interfacing residues, binding sites of proteins etc., while the non-structural evidences include, biological process, molecular function, cellular component and motif based similarities. These features are combined using Neural Network based classifier to predict true interactions. The algorithm showed encouraging results, when benchmarked for *Saccharomyces cerevisiae's* interactome, retrieved from the STRING database; with an accuracy of 92% for functional association networks, while on protein interaction networks the accuracy remained 83%.

INDEX TERMS Artificial intelligence, gene ontology, geometrical features, neural networks, P3I (prediction of protein-protein interactions), personalized medicine, protein binding sites, protein interaction network, protein structure.

I. INTRODUCTION

In a biological cell, proteins are the most abundant molecules after water. The mutual interactions of these molecules result into staggering biological complexities which ultimately is the foundation of life. With the advancements of next generation sequencing technologies, new genomes are sequenced, uncovering novel proteins, while their interactions remain unknown. An important step at the forefront of artificial intelligence and computational biology is to map the interactions of proteins that include: protein to DNA interactions, protein-protein interactions (PPIs), protein to RNA interactions and interactions of proteins with small molecules. A profound understanding of these interactions will lead to better development of antibacterial compounds, personalized therapies, and vaccines etc.

The classical approach to solve this problem is to use conventional wet lab experiments that include, "co-complex" interaction maps (which help to identify protein vs protein (bait) associations) [1], [2], Protein-fragment Complementation Assays (PCA) [3], or Yeast two-Hybrid (Y2H) [4] methods etc. These methods are inherently time consuming, and costly. Due to these limitations these methods can't be used to identify the true protein interaction networks of many unravelled genomes or even to prone the noisy interaction data present in various interaction databases.

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaochun Cheng.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

Many researchers in the last two decades utilized artificial intelligence based techniques incorporating a wide spectrum of biological information e.g., sequence homology, phylogenetic profiles, and protein's co-expression data, just to name a few, to predict genome-wide interactions of proteins [5]–[8]. However, recent comparative studies show that there is still a huge gap between the number of known genomes and their true interaction maps, [9], [10].

Most of the computational methods that achieve high prediction accuracy for PPIs, utilize structural information of proteins in one way or the other e.g., [11], [12]. The bottleneck for such methods is the sparsity of available structural information, which is evident from a vast gap between the number of protein sequences in sequence databases to the number of sequences with known structure. For example, a well known modal organism of *Saccharomyces Cerevisiae* has only 10% proteins with known structure [13].

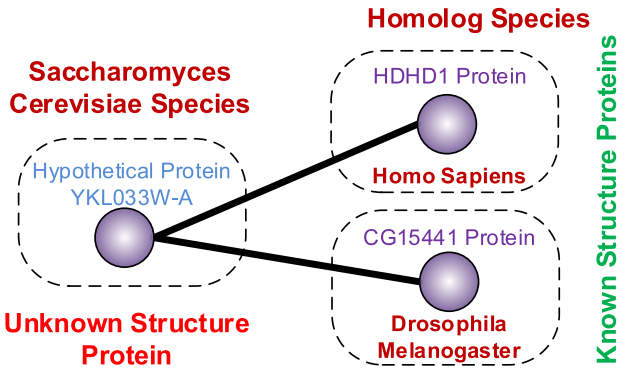


FIGURE 1. Sequence homology based connection between a functionally unknown protein of *Saccharomyces cerevisiae's* species with well known proteins of other species. [7].

The sparse structural proteins can be associated through structural homology both within and across species (an example can be seen in Figure 1), using the standardized structural databases, such as, the PDB (Protein Data Bank) database [14] etc. The structural homology links are established using shared geometrical features of individual templates. A method based on this type of information was proposed by [11], which has shown significant improvement in PPIs prediction accuracy. However, methods based on geometrical features exhibit degraded performance on proteins for which the homolog templates vary in structural details (i.e., the templates have varied geometry, interfaces, binding sites etc.). One way to deal with this issue is to incorporate additional information that could be non structural but strongly associates with the interactions, for example, the biological process ontology information of a protein which is believed to be linked with protein interactions.

In this paper, we propose a hybrid approach called *P3I (Prediction of Protein-Protein Interactions)* that integrates both structural and non-structural information for evaluating the likelihood of interaction given a pair of proteins. The novelty of *P3I* method lies in constructing a model that integrates the structural features with a number of non structural features such that the likelihood of true interaction is captured.

The structural features are mainly constructed by using the geometry of structural homolog templates, whereas, the non-structural features are designed using well known higher level knowledge that strongly associates to interactions, such as, the gene ontology based similarity, as well as the motif based similarity. The incorporated hybrid features are suitable for heterogeneous homolog templates because of their wider availability across various genomic sequences. The incorporated features are strongly associated with protein functions and since proteins interact to perform some functions, therefore; it significantly contributes in accurate elucidation of potential PPIs. Lastly, the hybrid features are passed to the artificial neural network classifier for evaluating the confidence of interaction for the input protein pair, which is a probability score depicting the confidence of interaction.

The remaining part of the paper is organized as follows: In Section 2, we outline the current state of the art techniques used to solve the problem of protein interaction prediction, starting from the most primitive feature based techniques to the most advanced feature based techniques (utilizing hybrid features). Furthermore, in Section 3, we present the proposed methodology (*P3I*) based on hybrid types of features to prediction protein interaction when given a pair of proteins as input. We demonstrate in Section 4, the effectiveness of our approach for predicting the interactions of the most widely studied model organism of *Saccharomyces cerevisiae*. In section 5, we provide a comprehensive discussion of the proposed technique on obtained results and their relevance to our hypothesis. In the last section we conclude our study along with potential future considerations.

II. BACKGROUND

The interactions of proteins are of paramount importance due to their involvement in almost all the biological phenomenon. With the advancement in sequencing technology, more specifically the emergence of the next generation sequencing technologies, molecular blueprints of many organisms are uncovered. However, the mutual association of these proteins i.e., their interaction information, still needs to be unravelled The precise knowledge of PPIs is restrained because of the intricate complexities of macro-molecules that tumble and twirl in a small droplet of water both within and outside a living cell. On the other hand, precise knowledge of PPIs is required to understand biological organisms as well engineered systems.

There are various experimental methods devised to predict protein-protein interactions. The most famous of these include: Protein-fragment Complementation Assays (PCA) [3], Yeast two-Hybrid (Y2H) [4], co-complex interaction maps [1], [2] etc. Unluckily, experimental techniques due to their biological nature have not been to scale up with the number of available genomes. Additionally, experimental approaches result in noise which ultimately results in either false negative or false positive interactions of proteins.

These limitations pose a challenge and only a few protein-protein interactions remain characterized. On the other hand, high throughput technologies are resulting into a huge volume of protein sequences who interaction remain uncharacterised

The very fundamental type of information that PPIs prediction algorithms integrate is the amino acid sequence (also called the primary sequence) of a protein. A protein's primary sequence is the most basic as well as amply available information which is widely used by PPI prediction algorithms It encodes many functional features of a protein. A lot of initial techniques in the area utilized this information to capture phylogeny using sequence and structural homology. A technique based on evolutionary information deduced from the structural conformations and protein activity relationships was proposed by [15]. This method utilizes phylogeny to estimate the possibility of interaction by integrating the phylogeny based features in their model.

Some scientists in [16], utilized the sequence information to make a cascaded classifier based on rotation forest and auto correlation descriptors. The authors of this study integrated the varied nature of two classifiers rotation forest and auto correlation descriptor to deal with the noisy nature of PPIs; which ultimately improved the prediction confidence of PPIs. The authors have shown enhancement in the interaction prediction results for the proteins of *Helicobacter pylori* and *Saccharomyces cerevisiae* species. Likewise, the authors in [17], have suggested to use the sequence information by computing alignments with different species proteins. In this work, they exploited the sequence similarity as the basis for deciphering potential interaction among proteins, as proteins interact to perform some functions and sequence similarity is strongly related to protein functions. For this purpose they chose Bayesian classifier to combine alignment based scores of potentially interacting proteins. A similar work was carried out by [18], which is also based on sequence only information. The authors reported significant progress over the previous works.

The next generation sequencing result in complete protein sequences of many modal organisms, that are utilized in many in-silico methods to cluster protein (by utilizing the interaction information) into related networks, thus providing understanding of functional associations for interacting proteins. A related work was done by [19], in which the researchers construct features from evolutionary trees and later combine them with some properties of the gene ontology. They reported significant improvement in terms of increase in true positive predictions when tested on the proteins of the *Saccharomyces cerevisiae* species.

In addition to utilizing the sequence information, various authors in [20], incorporate concepts from well known fields including, probability theory, graph theory, and graphical models etc., to predict the interactions among proteins. The authors proposed a probabilistic model to decipher the true interactions. They have reported predictions for human species proteins and precisely recovered 40,000 predictions. Furthermore, the high accuracy was attributed to probabilistic scores obtained by integrating interaction data, gene expression data as well as protein domain data.

In the same vein, many articles, e.g., [21], [22] etc., have focused on the utilization of probabilistic paradigms to predict protein-protein interactions. For example, in order to build an interaction prediction model the authors in [21] proposed a generative probabilistic model while using the graph biclique properties. They reported results for *Saccharomyces cerevisiae* species proteins. The researchers in this study converged to the hypothesis that the use of naïve duplication divergence model is more affective while comparing it with state of the art preferential attachment model. Moreover, a same kind of work has been reported by [22] that predicts the interaction network of *Saccharomyces cerevisiae* species proteins using probabilistic models. The features based on probabilistic models were combined with non-structural features using Bayesian classifier.

Protein domains is an important type of information linked with non-transient interactions. In addition to probabilistic models, some researchers e.g., [23], have used the protein domain knowledge to predict the protein-protein interactions. The initial work in this area has concentrated on utilizing features based on a single domain to function association. It was misunderstood that protein domains perform only one type of function. With the passage of time and evolution in technology, it became evident that a domain is associated with multiple types of protein activities. Similarly, the authors in [23], have used features based on multiple domains to predict protein-protein interactions.

An important type of information that is tightly bonded with protein functions, complex formations as well protein interactions is the structural information of a protein. The structural conformation of a protein determines the the potential interacting partners of that protein. A very famous technique was proposed by [24], in which the authors integrate features based on structure (as well as sequence) information to derive potential interactions. The researchers in this study, utilize sequence fingerprints and integrate them with structure based association scores to predict PPIs. Some researchers [25], suggested using protein structure data to identify the types of interactions a protein may be involved in. They utilized the statistics computed from structural geometry and used these score with SVM (Support Vector Machines) classifier to predict the PPIs. Another group of researchers in [26], incorporated another variation of geometrical features e.g., conserved patterns on interfaces of interacting proteins, to predict PPIs. This interaction specific structural information was reported to improve the prediction results.

Lastly, the most recent schemes that combine both structural and non structural information in a varied way are reported to outperform the previous methods, e.g., [13], [27]–[30]. The authors in these methods calculate the heterogeneous features and combine them using classical machine learning hypothesis functions e.g., Bayesian likelihood, SVM, neural networks (NN) etc. The authors in [13], reported

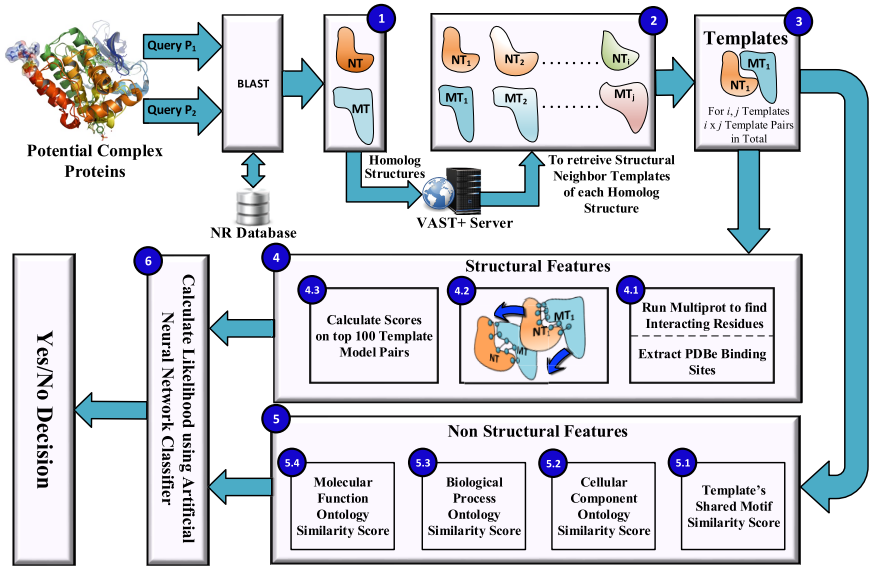


FIGURE 2. Flow diagram of the proposed model *P3I (Prediction of Protein-Protein Interactions)*.

that a combined information based model outperforms the single information based model. The method was validated for *Saccharomyces cerevisiae* species proteins.

In [27], the authors also compute hybrid features (i.e., structure based and non-structure based) and combine them using Bayesian likelihood. They reported to have higher accuracy based on hybrid feature than structural information alone. The hybrid features combine different aspect of the molecular activity of proteins, therefore; such information results in an enhanced confidence score of interactions, as interactions occur to perform some activity at the molecular level. In light of these observations, in our work we also employ hybrid features in a unique way that are known to be strongly related to the activities of proteins. We use them to build a model that predicts the PPIs with a high confidence.

III. MATERIAL & METHODS

The interaction network for the proposed algorithm was retrieved from the the most established *STRING* (Search Tool for the Retrieval of Interacting Genes/Proteins) database [31], covering direct (physical) as well as indirect (functional) interactions. Our method (*P3I*) combines both protein structural information e.g., interfacing residues on a protein's surface, potential binding residues etc., and non-structural information e.g., gene ontology based functional context, shared motifs etc., to infer protein-protein interactions. Another important aspect of this technique is that it exploits protein homology, and can easily be applied on a wide range of uncharacterised proteins resulting in increased coverage. We divide the proposed scheme into eight steps (as shown in Figure 2) and explain each step in detail, in the following subsections:

STEP 1: BLAST FOR HOMOLOG STRUCTURES

Our algorithm takes as input a potential protein pair, say P_1 and P_2 (we utilize *Uniprot* [32] IDs in our implementation). For each protein the prediction of PPIs becomes easier if we have some knowledge about its structure. In the first step of our algorithm, we associate each query protein P_i with some known structures. For this purpose, we utilize protein homology information to establish this link. We use homology because evolutionary relationships between species suggest that the homolog as well as orthologous proteins of different species, whose functions have been established before speciation event and which share high sequence similarity are more probable to have same structure as well as function.

A proteins amino acid sequence information is key to establish protein homology. It is well understood that if two proteins have an alignment score of more than 25 % then they are more likely to be homologs [33], [34] and are more likely to have similarity in structure and function. To find homolog similarity we run a BLAST [35] search for our input proteins against the protein's *NR* database, with an E-value cut-off of 0.0001. For each query protein we pick the model structure with the highest sequence similarity. The model structures are named as *NT* and *MT*. The model structure's detailed information is also downloaded from PDB [14], which will later be used in structural modeling.

STEP 2: FIND STRUCTURAL NEIGHBORS

Homolog structures alone are not sufficient for drawing any conclusion about interaction. In the second step in order to enhance prediction confidence, the method searches for structural neighbors of each model structures *NT*, and *MT*. For this purpose, the model structures are queried to VAST+ (Vector Alignment Search Tool Plus) service [36] to obtain

structural representatives. VAST+ service is provided by the NCBI (National Center for Biotechnology Information) consortium and is based on the standardized Molecular Modeling Database (MMDB) database. The VAST+ service returns structures that have very similar 3D conformation to the query structure. For each queried model structure, the service calculates geometrical similarities with MMDB database structures, without regard for sequence similarity. Thus the service is able to detect distant homolog structural neighbors. Among the returned structural representatives, we select the top ten structural representatives. The threshold of ten is selected to reduce computational overhead of the modeling phase (i.e., step 4 and 5). The identified structures are called: $NT_{\rm i}$ for model structure NT and $MT_{\rm i}$ for model structure MT, (where, $i = 1,2.....10$).

STEP 3: PAIRING OF REPRESENTATIVE TEMPLATES

For each query protein, so far, we have found one homolog structure and ten representative structures for each homolog structure. To evaluate the interaction strength of query proteins P_1 and P_2, we make pairs of all extracted structural neighbors i.e., $NT_{\rm i}$ with $MT_{\rm i}$ (where $i = 1,2.....10$), which means NT_1 pairs with MT_1, MT_2.... and so on up to MT_{10}, likewise we repeat pairing for NT_2, NT_3 .... up to NT_{10}. Thus, we generate a total of 100 templates. The template pairing is done to assess the encoded interaction strength of each template with the other. The higher this strength the more likely the proteins are to interact.

STEP 4: STRUCTURAL FEATURES

Once the template pairs are made, next step is to identify their interaction strength. To capture the physical interaction affinity, in the subsequent subsections, we calculate a number of structural features.

STEP 4.1) INTERACTING RESIDUES AND BINDING SITES

To identify the interacting residues and binding sites in the template pairs, we identify two types of residues in each template pair i.e., 1) interacting residues and 2) residues that are the binding sites, which will later be used in the model to calculate scores.

The first type of residues i.e., the interacting residues in the template pairs, are extracted using a tool called *Multiprot*, which is part of PRISM (PRotein Interactions by Structural Matching) protocol and is proven to be very successful in identifying interfacing residues [37], [38]. The Multiprot model is designed with the conception that globally different structure proteins can interact using small portions of similar conformation residues. By using this tool the interacting residues are computed for each template pair.

For identifying the second type of residues that are binding sites we utilize the PDBeMotif server by European Molecular Biology Laboratory (EMBL) [39]. PDBeMotif is a very rapid and powerful search tool that facilitates the exploration of binding sites in protein's structural templates. The binding sites are key to a substantial share of interactions that occur among proteins. The PDBeMotif is used to identify the binding sites, among the interacting residues of the template pairs.

STEP 4.2) INTERACTION MODELS

To capture overall interaction strength we build interaction models $Mod_{\rm ij}$ by overlaying the template pairs $NT_{\rm i}$ *and* MT_j over the homolog template NT and MT. Against each input query there are 100 models in total, built from 10x10 template pairs. Each model $Mod_{\rm ij}$ is used to calculate four structure based scores; in addition the scores of all the models are combined to give an overall structural score.

STEP 4.3) STRUCTURAL INTERACTION SCORES

The structural interaction scores are calculated by combining the structure information extracted for each template pair $NT_{\rm i}$ *and* MT_j i.e., # of interacting residues and the # of binding sites. This information is further utilized to calculate four scores for each interaction model $Mod_{\rm ij}$. Our first score is named as $\Re^{(1)}_{Mod_{ij}}$, where Mod_{ij} represents the interaction model. $\Re^{(1)}_{Mod_{ij}}$ denotes the number of shared interacting residues between the template pair $(NT_{\rm i}$ *and* $MT_j)$ and homolog structures $(NT$ and $MT)$, in interaction model Mod_{ij}. For arbitrary sequences of templates the $\Re^{(1)}_{Mod_{ij}}$ score integrates the interaction strength between the pairs by considering the interacting amino acids that are conserved.

$\Re^{(2)}_{Mod_{ij}}$ represents the second score and is evaluated by taking the ratio between the total number of interacting residues and the average number of interacting residues computed from the homolog templates (i.e., NT and MT). This score can be obtained using equation 1.

$$\Re^{(2)}_{Mod_{ij}} = \frac{\Re^{(1)}_{Mod_{ij}}}{Average(NT, MT)} \tag{1}$$

The third score $\Re^{(3)}_{Mod_{ij}}$ counts the potential binding sites in the shared interacting residues for interaction model Mod_{ij}. $\Re^{(3)}_{Mod_{ij}}$ score is calculated using equation 2.

$$\Re^{(3)}_{Mod_{ij}} = |\Re^{(1)}_{Mod_{ij}} \cap Binding_Sites(Mod_{ij})| \tag{2}$$

The final structural score $\Re^{(4)}_{Mod_{ij}}$ for interaction model Mod_{ij}, is calculated by taking shared binding sites in the homolog model pair and template pairs as shown in equation 3. $\Re^{(4)}_{Mod_{ij}}$ is the number of binding sites in the template that align to the number of binding sites in the model.

$$\Re^{(4)}_{Mod_{ij}} = |Binding_Sites(NT, MT) \cap Binding_Sites(Mod_{ij})| \tag{3}$$

When all scores are calculated for hundred interaction models Mod_{ij}, their effect is combined into one score namely, $\chi^{(k)}$ for each individual score $\Re^{(k)}_{Mod_{ij}}$ by taking the mean of as shown in equation 4.

$$\chi^{(k)} = \left(\frac{\sum_{i=1}^{10} \sum_{j=1}^{10} \Re^{(k)}_{Mod_{ij}}}{100} \right) \ldots For,\ k = \{1, 2, 3, 4\} \tag{4}$$

In addition to mean scores, we also take standard deviation scores of individual scores $\Re^{(k)}_{Mod_{ij}}$, to evaluate if the structural templates have mutual similarity or dissimilarity. If the differences among structural neighbors are too high, the standard deviation will be high; otherwise it will be low. The standard deviation scores $\chi^{(l)}$ for individual interaction model scores $\Re^{(k)}_{Mod_{ij}}$ are calculated using equation 5.

$$\chi^{(l)} = \left(\sqrt{\frac{\sum_{i=1}^{10} \sum_{j=1}^{10} \left(\Re^{(k)}_{Mod_{ij}} - \chi^{(k)} \right)^2}{100}} \right) \quad \ldots For,\ k = \{1, 2, 3, 4\}\ and\ l = \{5, 6, 7, 8\} \tag{5}$$

STEP 5: NON-STRUCTURAL FEATURES

In order to capture the biological context of interactions, we incorporate the non-structural features (or scores) that capture, on a higher level, the behavioural aspect of protein interaction networks. The first three non-structural scores are related to gene ontology whereas the last score is related to the fingerprints conserved in the interacting protein sequences.

Protein interactions occur under a collaborative objective and that objective is usually a function that they perform at the molecular level. Thus protein functions have strong correlation with protein interactions. For this purpose from the various classification schemes to standardize the definition of protein function, we select the state of the art Gene Ontology (GO) classification scheme [40]. The gene ontology is a widely used, structured, controlled vocabulary of protein activities where each activity in GO is called a term. GO terms provide wider coverage and consistency in annotating protein roles in the cellular context. Keeping in view the strong correlation between GO terms and protein interactions, we operate our first three non-structural scores based on gene ontology. More precisely each non-structural score is based on a sub-ontology of GO namely, cellular component, molecular function and biological process.

STEP 5.1) GO CELLULAR COMPONENT SIMILARITY SCORE

Cellular component ontology covers the localization aspect of a protein i.e., the parts of a cell or its extracellular environment where a protein localizes. Protein interactions are strongly correlated with the localization of proteins. The first non-structural feature namely $\chi^{(9)}$, is based on average cellular component ontology's semantic similarity among homolog templates. It is defined as:

$$\chi^{(9)} = \left(\frac{\sum_{i=1}^{10} \sum_{j=1}^{10} \frac{|GO_CC_{NTi} \cap GO_CC_{MTj}|}{min(|GO_CC_{NTi}|, |GO_CC_{MTj}|)}}{100} \right) \tag{6}$$

STEP 5.2) GO MOLECULAR FUNCTION SIMILARITY SCORE

The molecular function ontology covers the activities of proteins at the molecular level, such as binding or catalytic activities etc. Protein interact to perform one or more such activities, therefore molecular function similarity can be helpful in relating proteins for interaction. The second non-structural clue namely $\chi^{(10)}$, is based on average molecular function ontology's semantic similarity among homolog templates. It is defined as:

$$\chi^{(10)} = \left(\frac{\sum_{i=1}^{10} \sum_{j=1}^{10} \frac{|GO_MF_{NTi} \cap GO_MF_{MTj}|}{min(|GO_MF_{NTi}|, |GO_MF_{MTj}|)}}{100} \right) \tag{7}$$

STEP 5.3) GO BIOLOGICAL PROCESS SIMILARITY SCORE

Biological process ontology refers to sets of molecular events (with a definite start and end) which occur under the functional context of organized living units i.e., cells, tissues, organs, up to the complete organisms. To evaluate the event based similarity between homolog structures of interacting proteins, the third non-structural clue namely $\chi^{(11)}$, of our algorithm is defined as:

$$\chi^{(11)} = \left(\frac{\sum_{i=1}^{10} \sum_{j=1}^{10} \frac{|GO_BP_{NTi} \cap GO_BP_{MTj}|}{min(|GO_BP_{NTi}|, |GO_BP_{MTj}|)}}{100} \right) \tag{8}$$

STEP 5.4) TEMPLATE'S SHARED MOTIF SIMILARITY SCORE

A motif is a small conserved sequence present in a protein sequence having some functional significance. The number of motifs in sequences can also be used as a clue to determine its interaction tendency. Therefore, in addition to gene ontology based similarity, another type of non-structural feature that we incorporate is the shared motif similarity among templates. The shared motif among structural neighbor can give clue to their interaction. We utilize the motif information from the *PROSITE* database [41]. We calculate the shared motif based similarity score i.e., $\mathcal{F}^{(12)}$, using the following formula:

$$\chi^{(12)} = \left(\frac{\sum_{i=1}^{10} \sum_{j=1}^{10} \frac{|Motifs_{NTi} \cap Motifs_{MTj}|}{min(|Motifs_{NTi}|, |Motifs_{MTj}|)}}{100} \right) \tag{9}$$

STEP 6: PPI PREDICTION USING ARTIFICIAL NEURAL NETWORKS

In the last step, all structural and non-structural features are passed to an Artificial Neural Network (ANN) classifier for a decision about the input proteins P_1 and P_2, to be either interacting or non-interacting pair. The ANN is a practical technique for learning real valued, discrete valued, or vector valued hypothesis functions from the training data. One of the reason for choosing ANN classifier is its robustness to errors in the training data which has successfully been demonstrated in practical problems related to many different fields such as speech recognition, interpretation of visual scenes, learning robot control strategies, pattern recognition etc. On the other hand, probabilistic classifiers such as Bayesian, Probabilistic Graph Models etc., are also a good choice (as our features are mutually independent) but in some situations computing posterior probabilities becomes cumbersome, so that is why we chose the neural network classifier.

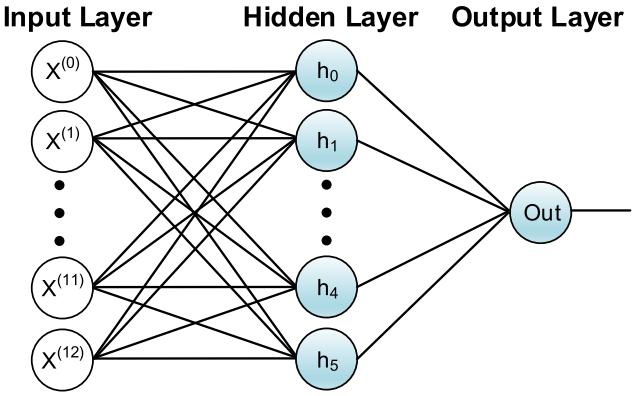


FIGURE 3. Artificial Neural Network architecture for PPI prediction.

The ANN's architecture that we employ is depicted in figure 3. We used three layer feed forward neural network with weights adjusted using stochastic gradient descent approach. The input layer of our architecture consists of twelve structural and non-structural features, namely $\chi^{(1)}, \chi^{(2)}, \chi^{(3)}, \ldots\ldots\ldots, \chi^{(12)}$ along with a bias input $\chi^{(0)}$. For training error, we use the binary cross entropy loss function, as shown in Equation 10.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}(y^{(i)} log h_\theta(x^{(i)}) + (1-y^{(i)})(1-log h_\theta(x^{(i)}))\right] \quad (10)$$

where, $x^{(i)}$ are the features and $y^{(i)}$ are the predicted labels. For the hidden layer we tried different possibilities to maximize generalization. The number of network parameters were adaptively adjusted by varying the number of hidden nodes. Each perceptron at hidden and output layer implements a sigmoid function. Among the various architectures tested, the network architecture (N_{hidden}=5, N_{output}=1) achieved the best performance with improved generalization. The output layer perceptron fires a higher value (reflecting interaction) if the input proteins P_1 and P_2 are found to be interacting otherwise it fires a lower value (depicting non-interacting protein pair).

IV. RESULTS

We benchmark our method (*P3I*) on the most widely used modal organism of *Saccharomyces cerevisiae* proteins. The interaction network was retrieved from the most established *STRING* (Search Tool for the Retrieval of Interacting Genes/Proteins) database [31], covering direct (physical) as well as indirect (functional) interactions. For each interacting pair in the dataset say *Pi* and *Pj*; our *P3I* method calculates twelve features (8 using structural information and 4 using non-structural information). These features are combined using Neural Network classifier to get a likelihood score of whether the input proteins *i.e., Pi* and *Pj*, interact or not.

A. STRING INTERACTION TYPES

The STRING database contains interactions for more than 2000 organisms, providing widest genome wide PPI coverage. In the STRING database, the *Saccharomyces cerevisiae* species' interactome consists of 59,394 interactions, as of August, 2017. The extracted interaction network includes proteins from many different species, with largest nodes from, *Saccharomyces cerevisiae, Arabidopsis thaliana, Homo sapiens, Rattus norvegicus, Arabidopsis thaliana,* and *Dictyostelium discoideum* species. We handle the false positive interactions by considering interactions with at least one experimental evidence. The filtered network contained 34,678 proteins.

B. EVALUATION MEASURES

We run our experiments in ten fold cross validation settings. We do random division of the interactome retrieved from STRING database into ten partitions. One is randomly selected for test while the model is trained on the remaining nine partitions. The process is repeated ten times. The input of our algorithm is a pair of proteins from the interactome (either interacting or non interacting). For each protein pair *Pi* and *Pj* in the interaction data-set, we consider the interaction of *Pi* and *Pj* to be unknown and then do prediction using our algorithm. For evaluation we compare the prediction results with the true interactions.

For evaluating the statistical strength of our algorithm we compute standard performance measures, such as: precision, recall, accuracy and F1 scores in the following manner:

$$precision = \frac{TP}{TP+FP}$$

$$recall = \frac{TP}{TP+FN}$$

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

and

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

where TP are the number of true positives, FP are the false positives, TN are the true negatives and FN are the number of false negatives.

C. UNPREDICTABLE FALSE NEGATIVES (UFN)

Before presenting the precision, recall, accuracy and F1 values. We explain the Unpredictable False Negative (UFN) terms that we exclude from our model. The UFN terms are false negative terms for which there is not enough biological evidence available, to give a prediction. To elaborate the concept of UFN terms it is pertinent to describe the way we utilize the concept of TP, TN, FP, and FN (both predictable and unpredictable FN) terms for our analysis. An illustration of the whole concept is presented in Figure 4. In this figure the true space represents all the original interactions in STRING dataset. The solution space on the other hand is the complete set of interactions that *P3I (Prediction of Protein-Protein Interactions)* takes into consideration for inclusion into prediction space; while prediction space is the set of all interactions predicted by our model *P3I*.

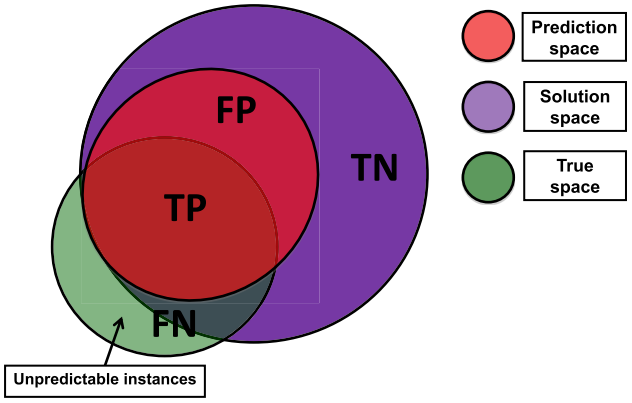


FIGURE 4. An illustration of TP, FP, TN, FN and Unpredictable FN for *Saccharomyces cerevisiae*'s interactome extracted from *STRING* database.

In the above stated spaces, the intersection between the true space and the prediction space is called the TP interactions; FP interactions are all interactions in prediction space excluding TP interactions; TN interactions are all the instances in solution space excluding the prediction space and true space instances. Lastly the FN instances on the other hand are those interactions which are present in the true space but are not present in the predicted space. These can broadly be categorized into UFN (not present in solution space) and predictable FN (present in solution space). The UFN interactions can't strictly be called as false negatives in the true sense because our method *P3I (Prediction of Protein-Protein Interactions)* didn't reject these interactions. The problem with UFN is the unavailability of related biological information that could be used to include proteins with these interactions into the solution space. Thus for our analysis we excluded the UFN terms from false negatives.

D. PRECISION, RECALL, ACCURACY AND F1 SCORES

In this section we present the precision, recall, accuracy and F1 score as shown in Table 1. The interaction network contains both the protein-protein interactions and functional associations of proteins. The protein interaction networks are inherently noisy and contain false positive interactions. To include, only the high confidence interactions, we filter the network into sets containing at least, 2 interaction evidences, 4 interaction evidences, 6 interaction evidences and more than six interaction evidences. As clearly can be seen that the network having at least six interaction evidences outperforms the other interactomes. The higher accuracy and precision values are at the cost of a slighter increase in the number of false negatives which slightly decreases the recall.

The proposed technique achieved an overall highest accuracy of 86.98%, precision of 86.89%, recall of 80.04% with an F1 score of 83.32%.

TABLE 1. Precision, Recall, Accuracy and F1 Scores for *Saccharomyces cerevisiae's* species with interactomes having 2, 4, 6 and 8 interaction evidences (containing both the protein-protein interactions and functional associations).

Number of interaction evidences	Precision	Recall	Accuracy	F1 Score
2	77.81%	88.41%	81.33%	82.77%
4	84.21%	86.15%	83.25%	85.17%
6	86.33%	81.36%	86.11%	83.77%
8	**86.89%**	**80.04%**	**86.98%**	**83.32%**

E. PRECISION AND ACCURACY OF INTERACTIONS VS FUNCTIONAL ASSOCIATIONS BASED EDGES

The STRING interaction database mainly contains two types of links among proteins 1) interaction links 2) functional association links. An interesting aspect of our results is to evaluate the false positives present in these edges. We filter the STRING interactome for yeast species with respect to the number of supporting evidences for each type of association i.e., 1) interaction links 2) functional association links. We made an overall eight interactomes for each type of association in the following manner: *Interactome 1* consisting of edges with at least one experimental evidence, *Interactome 2* consisting of edges with at least two experimental evidences and so on upto *Interactome 8* consisting of edges with at least eight experimental evidences.

To report the statistical bias of our classifier for each interactome i we evaluate the corresponding accuracy values. In addition, in order to capture the statistical variability of our predictions we also evaluate the precision values for each interactome i.

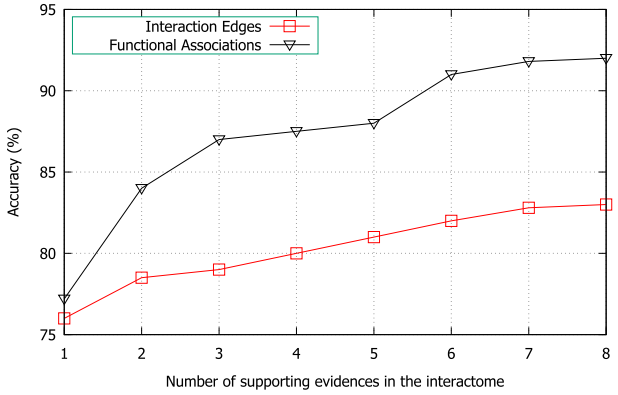


FIGURE 5. Accuracy values with multi-level interaction supporting evidences for *Saccharomyces cerevisiae*'s interactome extracted from *STRING* database.

The precision and accuracy values operated on each interactome i are reported in figure 6 and figure 5 respectively.

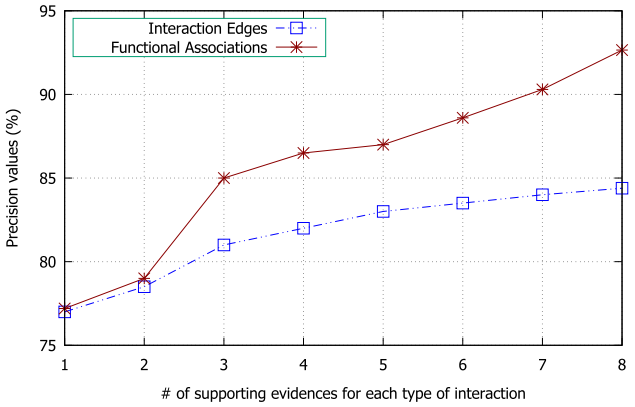


FIGURE 6. Precision values with multi-level interaction supporting evidences for *Saccharomyces cerevisiae*'s interactome extracted from *STRING* database.

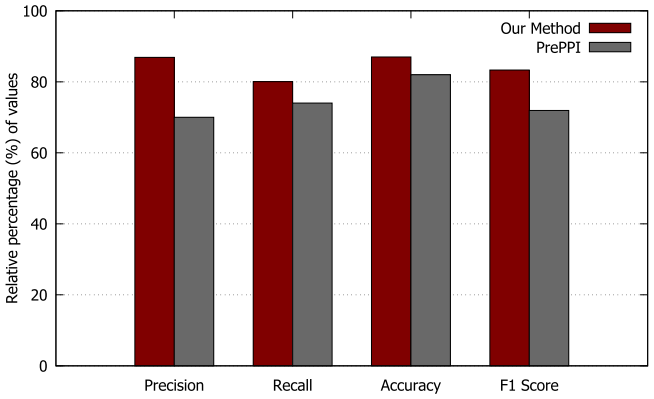


FIGURE 7. Comparison of Acurracy, Recall, Precision, and F1 score using our method (in maroon) and PrePPI (in gray) for interaction dataset of *Saccharomyces cerevisiae* proteins extracted from *STRING* database.

The model with eight experimental evidences achieved the overall best accuracy for both type of associations edges. Likewise, in figure 5, as expected, there is an increase in accuracy values with increasing number of supporting evidences for each type of edge. An important result that we conclude by looking at these curves is the higher accuracy of functional association edges in STRING database. From this we can conclude that functional association edges in STRING are much reliable (and have comparatively less noise) as compared with interaction type edges.

In Figure 6, we report the precision values with respect to the number of supporting evidences for each type of association in STRING database. Like accuracy, we see similar trend for precision. It is increasing as we increase the number of supporting evidences. In addition, functional association type edges have higher precision compared to interaction type edges. By looking at both the precision and accuracy curves, it can thus be concluded that functional association edges in STRING database are much reliable compared to interaction type edges.

We benchmark our results for best performing high confidence interaction network (consisting of proteins with eight number of experimental evidences per edge in the network) obtained from STRING database. Our model achieved an overall accuracy of 92% and precision of 92.65% when operated on functional associations type edges of STRING database. Likewise, it achieved an overall accuracy of 83% and precision of 84.39% when operated on interaction type edges of the STRING database.

F. COMPARISON WITH PrePPI APPROACH

In this section, we compare our algorithm with a recently proposed most famous technique called *Pre-PPI*, which was proposed by Zhang *et al.* [13]. In this paper, the author proposed to integrate different types of structural as well as non-structural biological data to infer protein-protein interactions. They further utilize Bayesian likelihood to calculate a probabilistic estimate on top of integrated structural and non-structural information. There is a plethora of techniques that use Bayesian reasoning in many different ways to precisely predict protein interactions. PrePPI uses Bayesian reasoning along with both structural and non-structural biological data and reported encouraging results; therefore, we selected PrePPI to compare with our approach.

To compare our method *P3I (Prediction of Protein-Protein Interactions)* with Pre-PPI we compute four statistical measures (i.e, precision, recall, accuracy and F1 score). In figure 7, the values for all four measures are reported for both techniques, when operated on *Saccharomyces cerevisiae's* interactome. It is important to mention here that the reported values are for high confidence interaction networks i.e., the interactome was filtered to retain interactions with at least eight experiments validating per interaction. From the figure it is evident that the proposed technique outperforms the PrePPI in all aspects i.e., precision, recall, accuracy and F1-score. Our method results in high precision because of the lower rate of false positives compared to PrePPI method. On the other hand higher accuracy is because of both true positives and true negatives in addition to lesser false positive rate. Another important measure of interest is the recall, which evaluates the number of true interactions within interactome that were missed by the classifier i.e., it not just considers true positive predictions but also utilizes false negative predictions (missed the classifier). For this measure as well, our techniques performs slightly better than PrePPI approach. The last measure is the F1 Score which is computed by combining recall and precision. The F1 score assesses the statistical strength of a classifier. The F1 score of *P3I (Prediction of Protein-Protein Interactions)* is also higher as both precision and accuracy values of our approach are higher than the prePPI approach (as shown in figure 7).

The outstanding performance of our technique can be attributed to a number of factors. The most important factor is the integration of PPI specific information i.e., protein binding sites, with non-structural features (which give global interaction context). Another important factor is the other

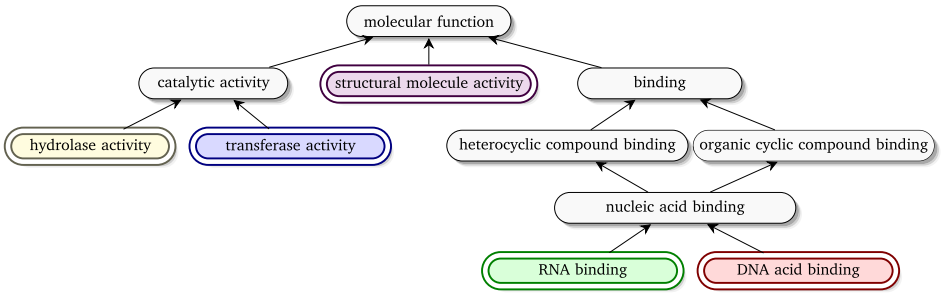


FIGURE 8. The five most frequent functions of *Saccharomyces cerevisiae's* proteins in the gene ontology.

type of structural information e.g., interfacing residues, interacting residues etc., which are also very helpful in exact modelling of the interaction behavior. Lastly, the non-structural clues e.g., motif similarity, and GO based different similarities provide the semantic information of interactions and are helpful in exact delineation of protein-protein interactions.

V. DISCUSSION

This section contains a brief analysis of how the proposed technique overcomes the identified problems. The analysis is carried out on five most frequent gene ontology terms present in the *Saccharomyces cerevisiae's* species.

A. WHY THE PPIs ARE DIFFICULT TO PREDICT?

Representing the precise behavior of protein-protein interactions in a model is a difficult task which can be attributed to many different factors. The most important factor effecting PPI prediction is the scale at which we want to characterize PPIs. Primarily, protein-protein interactions are grouped into two basic classes, namely: stable interactions and transient interactions. Whereas each type can have a degree of associated bonding strength.

Stable protein-protein interactions are unique and this uniqueness is because of their similarity to proteins that have been purified as multi-sub unit complexes. Stable PPI examples include haemoglobin and central RNA polymerase, the occurrence of apiece in a complex results in steady complexes (that are stable in their conformation).

The transient interactions, on the other hand, are mostly associated with biological processes frequently occurring in the cell. These type of interactions are temporary in nature and are the result of certain cellular conditions that govern the occurrence of these interactions. The example transient interactions include phosphorylation, localization to discrete zones of the cell and most frequently the conformational changes. While in contact with their interactors, momentary interacting proteins are included in a broad assortment of molecular level forms, including molecular transport, conformation, signaling, apoptosis, folding and cell cycling. Capturing these dynamics in a model is a daunting task and on top of that lack of associated biological data, makes it even more difficult for algorithms to maintain their robustness, while keeping a uniform prediction accuracy.

B. MOLECULAR FUNCTION BASED ANALYSIS OF PREDICTED PPIs

In this part we present the analysis of predicted PPIs with respect to most frequently occurring gene ontology terms of molecular function ontology. Proteins perform a multitude of functions at molecular level and a protein's function means molecular level, cellular level as well as localization aspects that a protein can be part of, including its interactions with other molecules (such as substrates, ligands, pathogens and other small compounds etc.). The PPIs occur to perform a variety of such functions. This analysis is important in delineating which interaction performs which type of molecular function. For this purpose we use the molecular function ontology of Gene Ontology (GO) classification scheme [40].

The GO is the current standard for protein functions due to desirable properties of this classification scheme, the important among those properties include, their coverage of many species, disjoint functional classes, format standardization etc. The gene ontology is a hierarchical arrangement of protein activities having Direct Acyclic Graph (DAG) organization.

We carried out our analysis on five most frequent gene ontology terms present in the *Saccharomyces cerevisiae's* species. The selected functions include hydrolase activity, RNA binding, DNA binding, transferase activity and structural molecule activity. In Baker's yeast species, the hydrolase activity has 665 sequence instances, transferase activity has 812, RNA binding has 872, DNA binding has 1251, and structural molecule activity with 289 sequence instance annotations. For broader coverage we selected only five ontology terms. It is important to note that these terms occur in high frequency and more than 85% proteins in the yeast species are annotated either directly or indirectly (through parent-child relationship) with these terms. The representation of selected proteins in gene ontology graph is shown in Figure 8.

The interactomes (containing both interactions and functional associations) are parsed for each ontology function and an overall accuracy is computed for each set of interactions. The accuracy values for each type of functional class are as

follows: 100% for DNA binding interactions, 98.5% for RNA binding interactions, 82.5% for hydrolase activity interactions, 84.3% for transferase activity interactions, and 77.2% for structural molecule activity interactions.

The outstanding prediction accuracy of our technique can be accredited to a number of reasons: First, the quality of interactions is very high as each interaction is supported by at least 8 experimentally verified interactions set. Secondly, as can be seen, the performance for DNA binding and RNA binding interactions outperforms the other three functional classes which is an indicator of the fact that the structural and non-structural information we integrate is more relevant to identifying binding type of interactions, which are also the most frequent type of interactions in yeast interactome.

VI. CONCLUSION

In this study, we present a novel approach for deciphering the interactions of proteins. The novelty of this approach lies in integrating the evidences in a hybrid fashion i.e., the structure based geometrical features, e.g., interacting residues, interfacing residues of proteins etc., with non-structural features, e.g., semantic similarity features based on gene ontology etc. The proposed approach is bench-marked on the *Baker's yeast* interactome extracted from the STRING database. The algorithm achieved higher accuracy on both types of STRING database networks i.e., protein interaction network as well as for functional association network of proteins. This indicates that the incorporated hybrid features are strongly linked with protein's functional activity. The proposed algorithm is modular in nature and can easily be extended to incorporte more evidences to further enhance the prediction confidence and applicability across species.

ACKNOWLEDGMENT

The authors would like to thank NUSyS Research Group (URL: http://pwr.nu.edu.pk/home/nusys/) for their support in running the scripts on their high computing platform.

REFERENCES

[1] G. Rigaut, A. Shevchenko, M. Wilm, M. Mann, B. Séraphin, and B. Rutz, "A generic protein purification method for protein complex characterization and proteome exploration," *Nature Biotechnol.*, vol. 17, no. 10, pp. 1030–1032, 1999.

[2] B. A. Shoemaker and A. R. Panchenko, "Deciphering protein–protein interactions. Part I. Experimental techniques and databases," *PLOS Comput. Biol.*, vol. 3, no. 3, p. e42, 2007.

[3] J. N. Pelletier, K. Arndt, A. Plückthun, and S. W. Michnick, "An *in vivo* library versus library selection of optimized protein protein interactions," *Nature Biotechnol.*, vol. 17, no. 7, pp. 683–690, 1999.

[4] T. Ito *et al.*, "A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*," *Proc. Nat. Acad. Sci. USA*, vol. 98, pp. 4569–4574, 2001.

[5] B. A. Shoemaker and A. R. Panchenko, "Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners," *PLOS Comput. Biol.*, vol. 3, no. 3, p. e43, 2007.

[6] L. Salwinski and D. Eisenberg, "Computational methods of analysis of protein protein interactions," *Curr. Opin. Struct. Biol.*, vol. 13, pp. 377–382, 2003.

[7] H. Ur Rehman, U. Zafar, A. Benso, and N. Islam, "A structure based approach for accurate prediction of protein interactions networks," in *Proc. Bioinf.*, 2016, pp. 237–244.

[8] A. Diaz, J. A. Dalton, and J. Giraldo, "Artificial intelligence: A novel approach for drug discovery," *Trends Pharmacol. Sci.*, vol. 40, no. 8, pp. 550–551, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165614719301361

[9] P. Braun, "An experimentally derived confidence score for binary protein-protein interactions," *Nature Methods*, vol. 6, pp. 91–97, 2009.

[10] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, "Protein interactions: Two methods for assessment of the reliability of high throughput observations.," *Mol. Cell. Proteomics*, vol. 1, no. 5, pp. 349–356, 2002.

[11] Q. C. Zhang, D. Petrey, R. Norel, and B. Honig, "Protein interface conservation across structure space," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 24, pp. 10896–10901, 2010.

[12] M. Wass, G. Fuentes, C. Pons, F. Pazos, and A. Valencia, "Towards the prediction of protein interaction partners using physical docking," *Mol. Syst. Biol.*, vol. 7, no. 1, p. 469, 2011.

[13] Q. C. Zhang and D. Petrey, "Structure based prediction of protein-protein interactions on a genome wide scale," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[14] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucl. Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.

[15] A. Valencia and F. Pazos, "Prediction of protein-protein interactions from evolutionary information," *Methods Biochem. Anal.*, vol. 44, pp. 411–426, Feb. 2003.

[16] J. F. Xia, K. Han, and D. S. Huang, "Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor," *Protein Peptide Lett.*, vol. 17, no. 1, pp. 137–145, Jan. 2010.

[17] L. Burger and E. V. Nimwegen, "Accurate prediction of protein protein interactions from sequence alignments using a Bayesian method," *Mol. Syst. Biol.*, vol. 4, p. 165, Feb. 2008.

[18] J. Shen, J. Zhang, X. Luo, W. Zhu, and K. Yu, "Predicting protein-protein interactions based only on sequences information," *Proc. Nat. Acad. Sci. USA*, vol. 104, pp. 4337–4341, Dec. 2006.

[19] J. Sun, J. Xu, Z. Liu, Q. Liu, and A. Zhao, "Refined phylogenetic profiles method for predicting protein-protein interactions," *Oxford J. Bioinf.*, vol. 21, no. 16, pp. 3409–3415, 2005.

[20] D. R. Rhodes, S. A. Tomlins, and S. Varambally, "Probabilistic model of the human protein-protein interaction network," *Nature Biotechnol.*, vol. 23, no. 8, pp. 951–959, 2005.

[21] R. Schweiger, M. Linial, and N. Linial, "Generative probabilistic models for protein–protein interaction networks—The biclique perspective," *Oxford J.*, vol. 27, no. 13, pp. i142–i148, 2011.

[22] M. S. Scott and G. J. Barton, "Probabilistic prediction and ranking of human protein-protein interactions," *BMC Bioinf.*, vol. 8, no. 1, p. 239, 2007.

[23] X. W. Chen and M. Liu, "Prediction of protein-protein interactions using random decision forest framework," *Oxford J. Bioinf.*, vol. 21, pp. 4394–4400, Oct. 2005.

[24] J. Espadaler, O. Romero, and R. M. Jackson, "Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships," *Oxford J.*, vol. 21, no. 16, pp. 3360–3368, Jun. 2005.

[25] M. Hue, M. Riffle, J. P. Vert, and W. S. Noble, "Large scale prediction of protein-protein interactions from structures," *BMC Bioinf.*, vol. 11, no. 1, p. 144, 2010.

[26] A. S. Aytuna, A. Gursoy, and O. Keskin, "Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces," *Oxford J. Bioinf.*, vol. 21 no. 12, pp. 2850–2855, Apr. 2005.

[27] H. Ur Rehman, I. Bari, A. Ali, and H. Mahmood, "A Bayesian approach for estimating protein-protein interactions by integrating structural and non-structural biological data," *Mol. BioSyst.*, vol. 13, no. 12, pp. 2592–2602, 2017.

[28] F. Rehman, O. Khalid, N. ul Haq, A. ur Rehman Khan, K. Bilal, and S. A. Madani, "Diet-right: A smart food recommendation system," *KSII Trans. Internet Inf. Syst.*, vol. 11, no. 6, pp. 2910–2925, 2017.

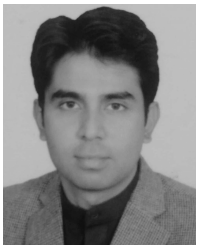
[29] X. Wang, R. Rak, A. Restificar, C. Nobata, C. Rupp, R. T. B. Batista-Navarro, R. Nawaz, and S. Ananiadou, "Detecting experimental techniques and selecting relevant documents for protein-protein interactions from biomedical literature," *BMC Bioinf.*, vol. 12, no. 8, p. S11, 2011.

[30] M. Shardlow and R. Nawaz, "Neural text simplification of clinical letters with a domain specific phrase table," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019.

[31] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, and M. Kuhn, "STRING v10: Protein–protein interaction networks, integrated over the tree of life," *Nucleic Acids Res.*, vol. 43, pp. D447–D452, Oct. 2015.
[32] T. U. Consortium, "UniProt: A hub for protein information," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D204–D212, 2015.
[33] A. Benso, S. Di Carlo, H. Ur Rehman, G. Politano, A. Savino, and P. Suravajhala, "A combined approach for genome wide protein function annotation/prediction," *Proteome Sci.*, vol. 11, no. S1, pp. 1–12, 2013.
[34] A. Mitrofanova, V. Pavlovic, and B. Mishra, "Prediction of protein functions with gene ontology and interspecies protein homology data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 3, pp. 775–784, May 2011.
[35] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990. [Online]. Available: http://blast.ncbi.nlm.nih.gov/Blast.cgi
[36] T. Madej, C. J. Lanczycki, D. Zhang, P. A. Thiessen, R. C. Geer, A. M. Bauer, and S. H. Bryant, "MMDB and VAST+: Tracking structural similarities between macromolecular complexes," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D297–D303, 2013.
[37] N. Tuncbag, A. Gursoy, R. Nussinov, and O. Keskin, "Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using prism," *Nature Protocols*, vol. 6, no. 9, pp. 1341–1354, 2011.
[38] M. Shatsky, R. Nussinov, and H. J. Wolfson, "A method for simultaneous alignment of multiple protein structures," *Proteins, Struct., Function, Bioinf.*, vol. 56, no. 1, pp. 143–156, 2004.
[39] A. Golovin and K. Henrick, "MSDmotif: Exploring protein sites and motifs," *BMC Bioinf.*, vol. 9, no. 1, p. 312, 2008.
[40] O. C. Gene, "Gene ontology consortium: Going forward," *Nucleic Acids Res.*, vol. 43, pp. D1049–D1056, 2015.
[41] N. Hulo and A. Bairoch, "The PROSITE database," *Nucl. Acids Res.*, vol. 34, pp. D227–230, Jan. 2006. [Online]. Available: http://prosite.expasy.org/

HAFEEZ UR REHMAN received the B.S. degree in computer science from the COMSATS Institute of Information Technology (CIIT), Abbottabad, in 2006, and the M.S. and Ph.D. degrees from the Politecnico di Torino University, Italy, in 2010 and 2014, respectively. He is currently an Associate Professor with the Department of Computer Science, FAST National University of Computer and Emerging Sciences—Peshawar, Peshawar. He has significant research contributions. He actively works in the fields of Bioinformatics and Medical Image Analysis. He has published a number of articles at many prestigious international research platforms.

USMAN HABIB received the master's degree from the Norwegian University of Science and Technology, (NTNU), Norway, in 2008, and the Ph.D. degree from the ICT Department, Technical University of Vienna, Austria. He is currently serving as an Assistant Professor and a Coordinator Graduate Program Committee with the Department of Computer Science, FAST National University of Computers and Emerging Sciences—Peshawar. He holds more than ten years of teaching and research experience spanning, since 2006. Along with teaching and research, he has also worked and successfully completed different industrial projects. He has been actively involved in research. He has authored several conferences and journal publications. His current research interests include machine learning, data analytics, fault detection, and diagnosis systems.



UMER IJAZ received the B.Sc. degree in electrical engineering, in 2007, and the M.S. degree in communication engineering and the Ph.D. degree in electronics and communication engineering from the Politecnico di Torino, Italy. After completing his graduation in B.Sc. degree, he worked with various national and multi-national companies. During his M.S./Ph.D. studies, he worked in collaboration with ST Microelectronics on various projects and conducted research with respect to Design and Development of X-direction 3D Avatar Animations Employing Wearable motion sensors. He is currently an Assistant Professor with the Electrical Engineering and Technology Department, Government College University, Faisalabad, Pakistan. He has published several research articles on various national and international forums. He is a HEC Approved Ph.D. Supervisor and a recipient of HEC-SRGP and IGNITE research grants to conduct various projects under his mentorship.

NAVEED ISLAM received the Ph.D. degree in computer science from the University of Montpellier II, France, in 2011. He is currently an Assistant Professor with the Department of Computer Science, Islamia College University, Peshawar, Pakistan. He has authored numerous articles in international journals and conferences. He is a Regular Reviewer of IEEE, Elsevier, and Springer Journals. His research interests include computer vision, machine learning, artificial intelligence, and data security.

ATTA UR REHMAN KHAN was the Director of the National Cybercrime Forensics Laboratory, Pakistan, the Head of Air University Cybersecurity Center, and the Conferences Chair of IEEE Islamabad Section. He is currently an Associate Professor and a Postgraduate Program Coordinator with the Faculty of Computing and Information Technology, Sohar University, Oman. His areas of research interests include cybersecurity, mobile cloud computing, ad hoc networks, and the IoT. He is a Steering Committee Member/Track Chair/Technical Program Committee (TPC) Member of over 60 international conferences. He also serves as a Domain Expert for multiple international research funding bodies. He has received multiple awards, fellowships, and research grants. He is serving as an Associate Editor for IEEE Access, the *Journal of Network and Computer Applications*—Elsevier; an Associate Technical Editor for the IEEE Communications Magazine; and an Editor for the *Journal of Cluster Computing*—Springer, *Oxford Computer Journal*, the IEEE SDN Newsletter, *KSII Transactions on Internet and Information Systems*, SpringerOpen *Human-centric Computing and Information Sciences*, *SpringerPlus*, and *Ad Hoc & Sensor Wireless Networks*—Journal.

RAHEEL NAWAZ is currently the Director of Digital Technology Solutions and Reader in analytics and digital education with Manchester Metropolitan University (MMU). He has founded and/or headed several research units specializing in artificial intelligence, digital transformations, data science, digital education, and apprenticeships in higher education. He has led on numerous funded research projects in the UK, EU, South Asia, and Middle East. He holds adjunct or honorary positions with several research, higher education, and policy organizations, both in the U.K. and overseas. He regularly makes media appearances and speaks on a range of topics, including Digital Technologies, Artificial Intelligence, Digital Literacy, and Higher Education. Before becoming a full-time academic, he served in various senior leadership positions for the private Higher and Further Education sector. He was an army officer before that.