# Detection of Protein Complexes Using a Protein Ranking Algorithm

Dr. Nazar Zaki* (Corresponding author)
Faculty of Information Technology, UAEU
Al Ain, P. O. Box 17551, UAE
Email: nzaki@uaeu.ac.ae

Dr. Jose Berengueres
Faculty of Information Technology, UAEU
Al Ain, P. O. Box 17551, UAE
Email: jose@uaeu.ac.ae

Dr. Dmitry Efimov
Faculty of Mechanics and Mathematics,
Moscow State University, Moscow, Russia
Email: diefimov@gmail.com

October 14, 2012

**Keywords:** PageRank algorithm, protein-protein interaction, essential protein, evolutionarily relationships, pairwise alignment.

**Abstract**

Detecting protein complexes from protein-protein interaction (PPI) network is becoming a difficult challenge in computational biology. There is ample evidence that many disease mechanisms involve protein complexes, and being able to predict these complexes is important to the characterization of the relevant disease for diagnostic and treatment purposes. This paper introduces a novel method for detecting protein complexes from PPI by using a protein ranking algorithm (ProRank). ProRank quantifies the importance of each protein based on the interaction structure and the evolutionarily relationships between proteins in the network. A novel way of identifying essential proteins which are known for their critical role in mediating cellular processes and constructing protein complexes is proposed and analyzed. We evaluate the performance of ProRank using two PPI networks on two reference sets of protein complexes created from MIPS, containing 81 and 162 known complexes respectively. We compare the performance of ProRank to some of the well known protein complex prediction methods (ClusterONE, CMC, CFinder, MCL, MCode and Core) in terms of precision and recall. We show that ProRank predicts more complexes correctly at a competitive level of precision and recall. The level of the accuracy achieved using ProRank in comparison to other recent methods for detecting protein complexes is a strong argument in favor of the proposed method.

Availability: Datasets, programs and results are available at http://faculty.uaeu.ac.ae/nzaki/ProRank.htm

# 1 Introduction

A fundamental challenge in human health is the identification of disease-causing genes. Many genetic diseases can be caused by multiple genes. Observations show that genes causing the same or similar diseases tend to lie close to one another in a network of protein-protein or functional interactions [1]. These genes are likely to be functionally related. Such functional relatedness could be exploited by measuring the evolutionarily relationships between genes to aid in the finding of novel protein complexes which eventually leads to the finding of disease genes. Direct PPI are one of the strongest manifestations of a functional relation between genes, so interacting proteins may lead to the same disease phenotype when mutated [2] and therefore predicting protein complexes is crucial. PPI networks possess valuable information regarding intra-cellular processes, and often involve interactions among protein complexes. Protein complexes are assemblies of proteins that have various functions inside the cell, such as cellular machines, rigid structures, dynamic signaling or metabolic networks and post-translational modification systems. Indeed, many diseases are caused directly by the inability for one component of a protein complex to bind correctly to the others. Examples abound in the literature of research into the detection and characterization of protein complexes in disease cases, and include general processes such as apoptosis [3], as well as specific diseases such as cancer, [4], HIV [5] and many others [6]. Therefore, the identification and characterization of the protein complexes involved is crucial to understanding the molecular events under normal and disease physiological conditions.

Several methods have recently been developed to predict protein complexes from the PPI. Protein complexes are commonly modeled as dense protein subnetworks. One of the most commonly used algorithms for predicting protein complexes via the dense protein subnetworks model is the molecular complex detection (MCode) algorithm [7]. Other clustering methods for inferring protein complexes include Markov clustering (MCL)[8], restricted neighborhood search clustering (RNSC) [9, 10], super paramagnetic clustering (SPC) and CFinder [11]. A comparative assessment [12] to evaluate the robustness of the above mentioned algorithms revealed that MCL is remarkably robust for graphing alterations. [13] recently developed an algorithm called Clustering-based on Maximal Cliques (CMC) to discover complexes from weighted PPI networks. They used an iterative scoring method to assign weight to protein pairs, which indicates the reliability of the interaction between proteins. CMC showed that the iterative scoring method can considerably improve the performance of CMC and other well known protein complex prediction methods. [14] utilized the core-attachment concept to develop an algorithm called Core to identify complexes from PPI network. The evaluation of the effectiveness of their proposed algorithm showed that Core can predict many more complexes and with higher accuracy than other cluster based methods. Nepusz et al. [15] has recently introduced protein clustering with overlapping neighborhood expansion. The method which they call it clusterONE detects potentially overlapping protein complexes from PPI data and showed better correspondence with reference complexes (from Munich Information Center for Protein Sequence (MIPS) catalog) than the results of other several popular methods. However, all of the above mentioned methods for predicting protein complexes have the following common limitations:

1. They are based on the idea of finding dense subgraphs however, they differ in the definition of a dense subgraph and the procedure to cluster the nodes into dense subgraphs. In order to achieve a breakthrough, we need a deeper understanding of the proteins within these complexes.

2. They are unable to detect complexes which contain few proteins or few interactions.

3. They have not incorporated biological knowledge such as structural or evolutionarily relationships between proteins within the complexes.

4. The datasets applied to these problems have been derived from high-throughput experiments, which, in the case of PPI, are known to have both high false-positive and high false-negative rates [16].

This paper introduces a novel method to detect protein-complexes from PPI by using protein ranking algorithm (ProRank). ProRank quantifies the importance of each protein based on the interaction structure of the protein network. This is somewhat analogous to PageRank algorithm [17, 18, 19, 20] at Google which is one of the most successful algorithms used to quantify and rank the importance of web pages. ProRank identifies essential proteins which play critical roles in mediating cellular processes, constructing complexes and adding potential therapeutic values. Several methods have been developed to identify highly connected protein nodes (hubs) [21, 22, 23] which could be very valuable in detecting protein complexes. However, hub protein prediction remains a very challenging task as properties that differentiate hubs from less connected proteins remain mostly uncharacterized. The PPI network often contains a significant number of highly connected proteins which are not genuine hubs. The network also includes a reasonable number of complexes which may be too small to constitute a hub (number of proteins in the complex $\leq 3$ proteins). Unlike hubs, essential proteins of these complexes may not be highly connected. To support this claim, we have analyzed two different yeast PPI datasets (PPI-D1) and (PPI-D2) (described in details in Section 3) and found that, among the top 25% highly connected proteins only 30% and 22%, respectively, are genuine hubs. This is motivated us to develop a novel way to identify and filter PPI network from which we call ”bridge”, ”shore” and ”fjord” proteins. These types of proteins ofen add noise to the PPI network.

Unlike the PageRank algorithm, ProRank, uses the evolutionarily relationships (in terms of similarity) to indicate the importance of the protein in the network. The essential protein is expected to interact and to have reasonably high similarity to most proteins within a complex. Sequence similarity often suggests evolutionary relationships between protein sequences that can be important for inferring similarity of structure or function [24]. The analysis on (PPI-D1) and (PPI-D2) datasets reveal that 79% and 84% of the complexes contain more than 50% proteins which share high similarity percentage ($\geq 60\%$), respectively.

To address the PPI network reliability problem, we used AdjustCD iterative method [25, 26, 27] to remove unreliable protein interaction pairs.

## 2 Method

The ProRank method consists of five steps:

- **Pruning:** Assigning weights to protein pairs and removing the unreliable interactions.

- **Filtering:** Analyzing the PPI network to distinguish between essential, bridge, shore and fjord proteins. These types of proteins add noise to the PPI dataset and they should be neglected.

- **Protein Similarity Calculating:** Calculating the similarity between proteins in the network.

- **Protein Ranking:** Ranking proteins based on the number of interactions and the similarity relationships.

- **Complex Detection:** Detecting protein complexes.

In the subsequent sections, we describe the above mentioned steps.

## 2.1 Pruning

To remove unreliable protein interaction pairs the AdjustCD iterative algorithm [25, 26, 27] is employed which is mainly based on the calculation of the number of common neighbors for protein pairs in the network. If two neighbor proteins are denoted as $u$ and $v$ then the CD-distance [26] between these proteins is defined as:

$$CD(u,v) = 1 - \frac{2|N_u \cap N_v|}{|N_u| + |N_v|}, \tag{1}$$

where $N_u$ and $N_v$ are the numbers of neighbors of proteins $u$ and $v$, respectively. Equation 1 was further modified by [27] to decrease the CD-distance for proteins with few number of interactions:

$$AdjustCD(u,v) = \frac{2|N_u \cap N_v|}{max(|N_u|, n_{avg}) + max(|N_v|, n_{avg})}, \tag{2}$$

where $n_{avg} = \frac{\sum_{x \in V} |N_x|}{n}$ is the average number of neighbors in the network, $n$ is the total number of nodes in the network.

Equations 1 and 2, clearly show how many 3-cliques are based on edges between proteins $u$ and $v$, but do not take into account the 3-cliques based on other edges from nodes $u$ and $v$. To solve this problem, [27] have suggested the iterative method which considers all 3-cliques based on edges from all neighbor nodes of $u$ and $v$:

$$w^k(u,v) = \frac{\sum_{x \in N_u \cap N_v} (w^{k-1}(x,u) + w^{k-1}(x,v))}{max(\sum_{x \in N_u} w^{k-1}(x,u), w_{avg}^{k-1}) + max(\sum_{x \in N_v} w^{k-1}(x,v), w_{avg}^{k-1})} \tag{3}$$

where $w^0(x,u) = 1$, if $x$ and $u$ interact, $w^0(x,u) = 0$, otherwise; $w_{avg}^{k-1} = \frac{\sum_{x \in V} \sum_{y \in N_x} w^{k-1}(x,y)}{n}$ is the average number of weights on $(k-1)^{th}$ step; $w^1(x,u) =$

$AdjustCD(x, u)$ and eventually $w^k(u, v)$ will show reliability of interaction between proteins $u$ and $v$.

## 2.2 Filtering

In this step bridge, shore and fjord proteins are identified. A bridge protein is a protein in the PPI network whose subgraph of neighbors is disconected. A fjord protein is a protein with neighbors which have a small number of connections between each other. A shore protein is a protein with at least one special neighbor which has significantly few connections to other proteins. The identification of these types of proteins will assist us in filtering the PPI network.

## 2.3 Protein similarity calculating

In this step, the similarity score between all proteins in the network is calculated. It is assumed that interacting protein pairs within the same complex may have a high similarity score [28]. A motivating observation is that the pairwise alignment score provides a relevant measure of similarity between protein sequences [29]. This similarity may incorporate biological information about the proteins' evolutionarily structural relationships [30]. The expectation is that proteins within the same complex may share evolutionarily structural relationships. The evolutionarily structural relationships could be reflected by the similarity scores calculated using Smith-Waterman algorithm as implemented in Fasta [31]. Smith-Waterman [32] has undergone two decades of empirical optimization in the field of bioinformatics and thus, considerable prior knowledge is implicitly incorporated into the pairwise sequence similarity scores and hence into the PPI network. The Smith-Waterman score $SW(v_0, v_1)$ between protein sequences $v_0$ and $v_1$ is the score of the best local alignment with gaps between the two protein sequences, computed by the SW dynamic programming algorithm [32].

All protein sequences in the PPI network are scored against each other. For instance, if we have a protein sequence $v$ then the corresponding score will be $F_v = f_{v_0}, ..., f_{v_n}$ where $n$ is the total number of the proteins in the network and $f_{v_i}$ is the $s_{v,i}$ score ($E$-value) between protein $v$ and the $i^{th}$ protein sequence. In this case, the default parameters are used; gap opening penalty and extension penalties of 11 and 1, respectively, and the BLOSUM62 matrix. We suppose $s_{v,i} = 0$ if there is no interaction exists between proteins $v$ and $i$. This process will result in generating the similarity matrix $S$ between proteins in the PPI network:

$$S = \begin{bmatrix} s_{1,1} & s_{1,2} & ... & s_{1,n} \\ s_{2,1} & s_{2,2} & ... & s_{2,n} \\ : & : & : & : \\ s_{n,1} & s_{n,2} & ... & s_{n,n} \end{bmatrix} \tag{4}$$

Furthermore, we create a new matrix $S^*$ which is basically the normalized matrix $S$ (by columns) such that $\sum_{i=1}^{n} s_{i,j} = 1$ for every $j = 1..n$.

## 2.4 Protein Ranking

Consider a network of $n$ proteins indexed by integers from 1 to $n$. This network is represented by the directed graph $G = (V, E)$. Here, $V := 1, 2, ..., n$ is the set of nodes (proteins) and $E$ is a set of edges (interactions) among the proteins.

The protein $i$ is connected to the protein $j$ by an edge, i.e., $(i, j) \in E$, if protein $i$ interacts with $j$.

The objective of the ProRank algorithm is to provide some measure of importance to each protein in the network. The ProRank value of protein $i \in V$ is a a real number in $[0, 1]$; we denote this by $x_i$.

The values of $x$ are ordered such that $x_i > x_j$ implies that protein $i$ is more important than protein $j$. The value of a protein is determined as a sum of the contributions from all proteins that have interactions to it. In particular, the value $x_i$ of protein $i$ is defined as:

$$x_i = \sum_{j \in \tau_i} s^*_{i,j} \cdot x_j \tag{5}$$

where $\tau_i := \{j : (j, i) \in E\}$, i.e., this is the index set of proteins interacting to protein $i$, and $n_j$ is the number of outgoing links of protein $j$. It is customary to normalize the total of all values so that $\sum_{i=1}^{n} x_i = 1$

Let the values of $x$ be in the vector form where $x \in [1, 0]^n$. Then, from Equation 5, the PageRank algorithm can be rewritten as:

$$x = S^* x, x \in [1, 0]^n, \sum_{i=1}^{n} x_i = 1 \tag{6}$$

where $S^*$ is the normalized similarity matrix calculated in section 2.3.

Note that the vector $x$ is a nonnegative eigenvector corresponding to the eigenvalue 1 of the nonnegative matrix $S^*$. In general, for this eigenvector to exist and to be unique, it is critical that the PPI network is strongly connected. To find the eigenvector corresponding to the eigenvalue 1 a modified version of the values has been introduced in [17] as follows:

Let $m$ be a parameter such that $m \in (0, 1)$, and let the modified interaction matrix $M \in \Re^{n \times n}$ be defined by:

$$M := (1 - m) S^* + \frac{m}{n} \mathbf{1} \tag{7}$$

where $\mathbf{1}$ is a $n \times n$ matrix with all elements equal to 1.

In the original algorithm in [17], a typical value for $m$ is reported being $m = 0.15$; The value of $m$ is too small and it does not really influence the results. It is rather important in avoiding convergence problems. Notice that $M$ is a positive stochastic matrix. Thus, according to Perron theorem [33], this matrix is primitive; in particular, $|\lambda| = 1$ is the unique maximum eigenvalue. To find corresponding eigenvector $x$ we apply the following formula:

$$x(k+1) = Mx(k) = (1-m)S^{*}x(k) + \frac{m}{n}\mathbf{1}, \tag{8}$$

where $x(k) \in \Re^{n\times 1}$ and the initial vector $x(0) \in \Re^{n\times 1}$ is a probability vector. Now let us expand on the convergence rate of this scheme. Let $\lambda_1(M)$ and $\lambda_2(M)$ be the largest and the second largest eigenvalues of $M$ in magnitude. Then, for the power method applied to $M$, the asymptotic rate of convergence is exponential and depends on the ratio $|\lambda_2(M)/\lambda_1(M)|$. Since $M$ is a positive stochastic matrix, we have $\lambda_1(M) = 1$ and $|\lambda_2(M)| < 1$. Furthermore, it is shown in [19] that the structure of the link matrix $M$ leads us to the bound

$$|\lambda_2(M)| \leq 1 - m. \tag{9}$$

Figure 1 shows a simplified network with four proteins that illustrates the ranking step.

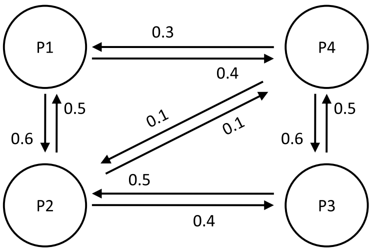

Figure 1: A PPI network with four proteins.

The link matrix $S^{*}$ and the modified link matrix $M$ can easily be constructed by (2) and (3), respectively, as:

$$S^* = \begin{bmatrix} 0 & 0.5 & 0 & 0.3 \\ 0.6 & 0 & 0.5 & 0.1 \\ 0 & 0.4 & 0 & 0.6 \\ 0.4 & 0.1 & 0.5 & 0 \end{bmatrix}, M = \begin{bmatrix} 0.038 & 0.463 & 0.038 & 0.293 \\ 0.548 & 0.038 & 0.463 & 0.123 \\ 0.038 & 0.378 & 0.038 & 0.548 \\ 0.378 & 0.123 & 0.463 & 0.038 \end{bmatrix} \tag{10}$$

Using the power method, the eigenvector $x$ can be computed as $x = [0.214, 0.283, 0.259, 0.244]^T$ with $m = 0.15$. Notice that the protein (P2) has the largest value since it interacts with three other proteins. On the other hand, protein (P2) and (P4) appear to have a similar number of interactions. However, P2 yielded a larger value since the similarities to other proteins (P1 and P3) are higher. This indicates that, similarity evidence has a major effect in ranking proteins in PPI network.

## 2.5 Complex Detection

Once all proteins in the network are ranked, we used the spoke model (the bait is predicted to interact with all members of a complex) to identify the protein complexes. Most of the previous methods have used the matrix model (all members of the complex are predicted to interact with all other members in the complex) to identify protein complexes [7]. Once the essential protein is identified, we can simply detect all neighbors connected to it. Following the complexes identification, we check all predicted complexes. If more than 50% of the neighbors of each proteins from complex $C1$ are in complex $C2$ we merge the two complexes.

## 2.6 How ProRank works

Figure 2 illustrates how ProRank works. Let us consider a hypothetical PPI network of 19 proteins and 5 complexes ($C1, C2, .., C5$).

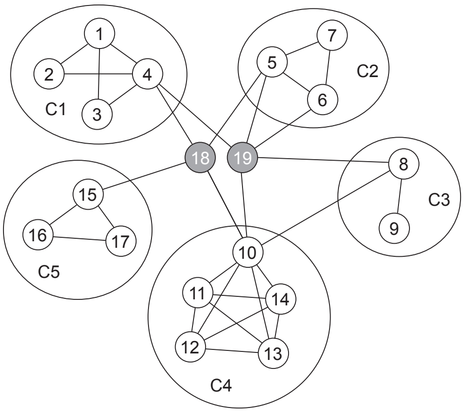

Figure 2: A hypothetical PPI network.

Complexes such as $C3$ and $C5$ are included to show that ProRank is capable of detecting complexes which consists of few proteins. Complex $C5$ contains 3 proteins with only 2 interactions (no 3-cliques). Two proteins 18 and 19 are included to demonstrate noise in the PPI network. Neither of these two proteins should be part of any predicted complex even though they connect to many proteins from various complexes. These two proteins could easily be confused as hub proteins.

For simplicity we assume that the interactions between proteins in the network are reliable and that the similarities between them are uniform. We first start by analyzing the network to identify bridge, fjord and shore proteins. Figure 3 and Figure 4 show some examples of bridge and fjord proteins identified. No shore protein is detected in this case. Protein 18 for instance, is identified as bridge type protein since it is connected to a disconnected subgraph of neighbors. No connection exits between proteins $4, 5, 15$ nor 10. On the other hand, protein 19 is identified to be ”fjord” protein since it is connected to neighbors which have a small number of connections between each other (protein 5 is connected to 6, 6 to 8 and 8 to 10).

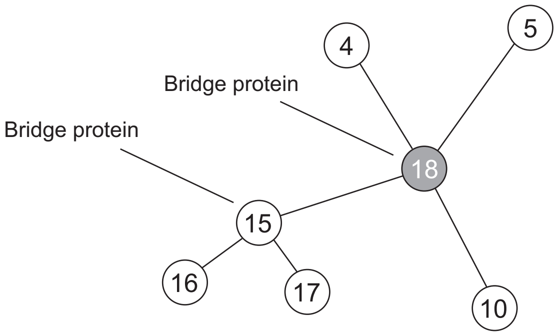

Figure 3: Example of bridge proteins.

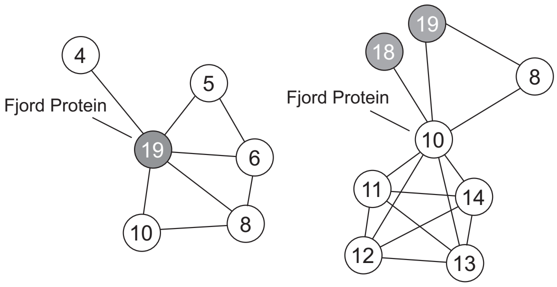

Figure 4: Example of fjord proteins.

In Table 1 we listed each protein with its corresponding complex, rank score and type. It is no surprise to see that protein 10 ranks the highest because it is directly connected to 7 proteins in the network. Proteins $10, 4, 19, 8, 5$ and 6 are identified as fjord proteins. Two proteins 18 and 15 are identified as bridge proteins. The rest are classified as essential proteins. ProRank then selects the highly ranked essential protein 14 and use the spoke model to pull it’s neighbors proteins $(10, 11, 12$ and $13)$ which forms complex $C4$. Once a protein is assigned to a complex it will not be considered further as essential type of protein. Next, ProRank selects the next protein: protein 1. It will pull the neighboring proteins

$(2, 3$ and $4)$. This process continues till complexes $C1, C2, C3$ and $C4$ are all detected. However, proteins 16 and 17 will be able to pull the same protein 15 and therefore, two complexes will be detected which contain proteins (17 and 15), and (16 and 15) respectively. This incongruence will be solved by the merging step as mentioned in Section 2.5.

Table 1: Analyzing the hypothetical network and ranking/identifying the types of proteins in the network.

| Protein | Cmplx | Ranking Score | Type of protein |
|---|---|---|---|
| 10 | $C4$ | 0.0956 | Fjord |
| 4 | $C1$ | 0.0799 | Fjord |
| 19 | $-$ | 0.0734 | Fjord |
| 18 | $-$ | 0.0649 | Bridge |
| 15 | $C5$ | 0.0638 | Bridge |
| 8 | $C3$ | 0.0617 | Fjord |
| 5 | $C2$ | 0.0616 | Fjord |
| 6 | $C2$ | 0.0610 | Fjord |
| 14 | $C4$ | 0.0540 | Essential |
| 13 | $C4$ | 0.0540 | Essential |
| 12 | $C4$ | 0.0540 | Essential |
| 11 | $C4$ | 0.0540 | Essential |
| 1 | $C1$ | 0.0511 | Essential |
| 3 | $C1$ | 0.0350 | Essential |
| 2 | $C1$ | 0.0350 | Essential |
| 7 | $C2$ | 0.0329 | Essential |
| 17 | $C5$ | 0.0244 | Essential |
| 16 | $C5$ | 0.0244 | Essential |
| 9 | $C3$ | 0.0192 | Essential |

## 2.7 Evaluation measures

To evaluate the accuracy of ProRank, the Jaccard index measure is introduced and it is defined as follows:

$$Acc(K, P) = \frac{|K \cap P|}{|K \cup P|}, \tag{11}$$

where $K$ and $P$ are the known and predicted complexes respectively, and $|.|$ is the number of proteins. The complex $P$ is defined to match $K$ if $Acc(K, P) \geq 0.5$.

To estimate the cumulative quality of the prediction, we compare the number of matching complexes with the number of known complexes using recall ($Rec_c = \frac{N_{MK}}{N_K}$) and precision ($Prec_c = \frac{N_{MP}}{N_P}$), where $N_{MK}$ is a number of matching known complexes, $N_{MP}$ is a number of predicted known complexes, $N_K$ is a number of known complexes and $N_P$ is the number of predicted complexes.

In some cases (when predicted complexes vary greatly in size) it is necessary to consider the quantity of proteins in the known and predicted complexes for prediction accuracy estimation. As before two characteristics are used recall:

$$Rec_N = \frac{\sum_{i=1}^{N_{MK}} |C_i|}{\sum_{i=1}^{N_K} |K_i|}, where|C_i| = max_{P_j:Acc(K_i,P_j)\geq 0.5}|K_i \cap P_j| \quad (12)$$

and precision,

$$Prec_N = \frac{\sum_{i=1}^{N_{MP}} |C_i|}{\sum_{i=1}^{N_P} |P_i|}, where|C_i| = max_{K_j:Acc(P_i,K_j)\geq 0.5}|P_i \cap K_j| \quad (13)$$

Calculations were made of precision and recall at complex and complex-protein levels.

# 3 Experimental work and results

Yeast has long been known as a highly effective model organism for mammalian biological functions and diseases [34]. We evaluated the effectiveness of ProRank using two different yeast PPI datasets. The first dataset (PPI-D1) was prepared by Gavin et al. [35]. The second dataset (PPI-D2) is a combined PPI dataset containing yeast protein interactions generated by six individual experiments, including interactions characterized by mass spectrometry technique [36, 37, 35, 38], and interactions produced using two-hybrid techniques [39, 40]. Two reference sets of protein complexes are used in these experiments. The first set of complexes (Cmplx-D1) is created from MIPS [41]. In the case of MIPS, only complexes that were manually annotated from DIP interaction data are considered. Following Leung et al. [14], complexes of sizes less than 5 proteins are excluded and therefore, 81 complexes are considered. The second set of complexes (Cmplx-D2) comprises of 162 hand-curated complexes (size no less than 4 proteins) from MIPS [42]. This dataset was used by Liu et al. [13] to evaluate the performance of the CMC method.

Employed first was the AdjustCD method [25, 26, 27] to assign weights to the PPI protein pairs in PPI-D1 and PPI-D2 to determine reliable interactions. The influence of using only reliable PPI interaction pairs is shown in Table 2.

Table 2: Influence of using reliable PPI interaction pairs.

| | Cmplx-D1 | | | Cmplx-D2 | | |
|---|---|---|---|---|---|---|
| Threshold | Matched Cmplx | $Rec_c$ | $Prec_c$ | Matched Cmplx | $Rec_c$ | $Prec_c$ |
| 0 | 56 | 0.691 | 0.421 | 48 | 0.2963 | 0.195 |
| 0.05 | 54 | 0.667 | 0.429 | 49 | 0.302 | 0.219 |
| 0.1 | 56 | 0.691 | 0.452 | 54 | 0.333 | 0.28 |
| 0.125 | **57*** | 0.704 | 0.496 | 53 | 0.327 | 0.294 |
| 0.15 | 54 | 0.667 | 0.514 | **59*** | 0.364 | 0.353 |
| 0.2 | 55 | 0.679 | 0.557 | 55 | 0.34 | 0.404 |
| 0.25 | 52 | 0.642 | 0.584 | 51 | 0.315 | 0.415 |
| 0.3 | 50 | 0.617 | 0.595 | 49 | 0.302 | 0.402 |

The table shows that better results are achieved by using threshold values of 0.125 and 0.15 for PPI-D1 and PPI-D2, respectively. Pairs with weight below the above mentioned threshold values are eliminated from the two PPI networks. Table 3 shows details of the two different PPI datasets and the number of proteins and interactions before and after the pruning step.

Table 3: Details of the PPI datasets before and after the pruning step.

| | Before pruning | | After pruning | |
|---|---|---|---|---|
| Datasets | No. of int. | No. of prot. | No. of int. | No. of prot. |
| PPI-D1 | 6,531 | 1,430 | 4,687 | 990 |
| PPI-D2 | 23,399 | 3,869 | 6,993 | 1,443 |

Following the pruning step, we analyzed PPI-D1 and PPI-D2 to identify and filter the bridge, fjord and shore protein types. Table 4 shows the distribution of essential, shore, fjord and bridge proteins identified in PPI-D1 and PPI-D2. It also shows the distribution of the top 25% proteins.

Table 4: Number of essential, shore, fjord and bridge proteins in PPI-D1 and PPI-D2.

| Dataset | Essential | Shore | Fjord | Bridge |
|---|---|---|---|---|
| PPI-D1 (top 25%) | 108 | 124 | 121 | 7 |
| PPI-D1 (all) | 605 | 177 | 191 | 17 |
| PPI-D2 (top 25%) | 78 | 137 | 138 | 7 |
| PPI-D2 (all) | 853 | 230 | 331 | 29 |

It is noteworthy to mention that 30% and %22 of the top highly ranked 25% proteins in PPI-D1 and PPI-D2, respectively, are essential proteins.

Table 5 shows the positive effect on performance by filtering using shore, fjord and bridge protein labeling,

Table 5: Effect of filtering on ProRank performance.

| | With Filtering | | | Without Filtering | | |
|---|---|---|---|---|---|---|
| Dataset | Matched Cmplx | $Rec_c$ | $Prec_c$ | Matched Cmplx | $Rec_c$ | $Prec_c$ |
| Cmplx-D1 | 57 | 0.704 | 0.496 | 45 | 0.577 | 0.446 |
| Cmplx-D2 | 59 | 0.364 | 0.353 | 35 | 0.327 | 0.238 |

Following the filtering step, the similarity matrix $S$ is calculated. In this case, the default parameters are used; gap opening penalty and extension penalties of 11 and 1, respectively, and the scoring matrix BLOSUM62.

To analyze the effect of using similarity scores, we measure the similarity between proteins within the same complexes. The analysis on PPI-D1 and PPI-D2 datasets reveal that 79% and 84% of the complexes contain more than 50% of similar proteins. We consider that two proteins are highly similar if their similarity score is greater or equal to 60%. Four extra complexes were predicted correctly when the similarity matrix is used in place of the uniform similarity matrix.

Table 6 shows number of predicted complexes for different precision thresholds $t = \{0.5, 0.6, 0.7, 0.8, 0.9\}$, where $t$ is the number of proteins found in the predicted complex that are also found in the known complex, divided by the size of the predicted complex.

Table 6: Matching complexes for $t = \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

| Dataset | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | Cmplx Generated |
|---|---|---|---|---|---|---|
| Cmplx-D1 | 57 | 50 | 41 | 29 | 20 | 115 |
| Cmplx-D2 | 59 | 47 | 33 | 18 | 9 | 167 |

Table 7 shows complex detection accuracy in terms of $Rec_c$, $Prec_c$, $Rec_N$ and $Prec_N$ ($t = 0.5$).

Table 7: Complex detection accuracy in terms of $Rec_c$, $Prec_c$, $Rec_N$ and $Prec_N$ ($t = 0.6$).

| Dataset | $Rec_c$ | $Prec_c$ | $Rec_N$ | $Prec_N$ |
|---|---|---|---|---|
| Cmplx-D1 | 0.716 | 0.504 | 0.647 | 0.563 |
| Cmplx-D2 | 0.364 | 0.353 | 0.238 | 0.279 |

## 3.1 Comparison to the existing methods

This section compares the performance of ProRank to other methods for detecting protein complexes. The comparison was conducted based on Cmplx-D1 and Cmplx-D2. In Table 8 ProRank is compared to CMC, CFinder, MCL, ClusterONE and MCode based on Cmplx-D1. All the mentioned methods were applied after the pruning step. The iterative scoring parameter $k$ is set to 2 as it was shown by Liu et al. [13] that the iterative scoring method reaches the best performance when $k = 2$. For MCL, inflation is set to 1.8. For MCode, the depth is set to 100, node score percentage to 0, and percentage for complex fluffing to 0.2 as suggested by Brohee et al. [43]. For CFinder, we set $k$-clique size to 4. The rest of the parameters are set to their default values. Details of the experimental work and parameters setup of the mentioned methods are available from [13].

As shown in Table 8, ProRank is able to detect more matched complexes (59 matching complexes) than any of the published results [13] of other state-of-the-art methods with higher recall and precision.

Table 8: Performance comparison of ProRank to CMC [13], CFinder [11], MCL [8] and MCode [7].

| Methods | Detected Complexes ($t = 0.5$) | $Rec_c$ | $Prec_c$ |
|---|---|---|---|
| ProRank | 59 | 0.364 | 0.353 |
| CMC | 56 | 0.346 | 0.297 |
| CFinder | 46 | 0.284 | 0.379 |
| MCL | 51 | 0.315 | 0.353 |
| MCode | 39 | 0.241 | 0.330 |

For general purposes ProRank is compared to Core [14], MCL [8], Mcode [7], ClusterONE [15] and Cfinder ($k$-clique size of 3 and 4) [11] based on Cmplx-D1. For ClusterONE and since it is unweighted PPI, we set the density threshold to 0.5, the merging threshold to 0.8 and the penalty value of each node to 2.

Table 9 shows the ability of ProRank to detect significantly more complexes than the other methods.

Table 9: Comparison of ProRank to Core [14], MCL [8], Mcode [7], ClusterONE [15] and Cfinder ($k$-clique size = 3 and 4) [11] based on PPI-D1 and Cmplx-D1.

| Methods | Detected Complexes ($t = 0.6$) | $Rec_c$ | $Prec_c$ |
|---|---|---|---|
| ProRank | 50 | 0.617 | 0.435 |
| Core | 35 | 0.4321 | 0.1510 |
| MCL | 32 | 0.3951 | 0.1380 |
| Mcode | 23 | 0.2840 | 0.2191 |
| ClusterONE | 48 | 0.593 | 0.318 |
| Cfinder ($k$-clique size = 3) | 22 | 0.2716 | 0.2245 |
| Cfinder ($k$-clique size = 4) | 25 | 0.3086 | 0.3521 |

# 4 Conclusion and Discussion

In this paper, we introduce a novel method called ProRank used for detecting protein complexes from a PPI network of yeast. The level of accuracy achieved using ProRank for detecting protein complexes is a strong argument in favor of this proposed method. The improvement exhibited by ProRank in detecting protein complexes is due to three main reasons. Firstly, the utilization of a powerful method which is based on a PageRank algorithm. Secondly, the incorporation of the evolutionary relationships between proteins in the PPI network. Thirdly, utilizing robust methods to analyze the topology of the network which assists in the remove of noise (filtering) and unreliable interacting protein pairs (pruning) from the network. Therefore, ProRank has a great potential to identify novel complexes.

Furthermore, a novel way of identifying essential proteins is proposed and analyzed. Finding essential protein could lead to the understanding of the protein complex structure and disease mechanism as they could be good targets for antimicrobial agents. It could also be a key to study the centrality-lethality rule phenomenon [44], as well as processes such as apoptosis [3] and other disease-related mechanisms. In this case, the results based on PPI-D1 dataset revealed that proteins such as “YGR090W”, “ER172C”and “YDL055C”are the highly ranked proteins in the network. YGR090W involves in nucleolar processing of pre-18S ribosomal RNA and ribosome assembly, YER172C require for disruption of U4/U6 base-pairing in native snRNPs to activate the spliceosome for catalysis; homologous to human U5-200kD [SGD] and YDL055C require for normal cell wall structure and therefore all are important proteins to study. ProRank was also able to accurately predict several protein complexes such as oligomeric Golgi (COG), Transcription Factor IIA tau and the Origin Recognition complex (ORC). COG plays a key role in the intracellular trafficking, processing and secretion of glycoproteins, glycolipids and proteoglycans [45]. Defects in conserved oligomeric Golgi complex could result in multiple deficiencies in protein glycosylation. On the other hand, Transcription Factor IIA tau is associated with undifferentiated cells and its gene expression [46] while ORC plays a role in the establishment of silencing at the mating-type loci Hidden MAT Left (HML) and Hidden MAT Right (HMR).

The experimental comparisons have shown that ProRank is better than or competitive to other existing methods in many aspects. Unlike the existing methods, ProRank focuses on identifying important proteins with great influence in the PPI network. As shown in Figure 5, ProRank is able to identify the protein YPR029C as an essential protein with a high ranking score. Based on this information other proteins in the complex (highlighted in black) can easily be identified such as YLR170C, YPL259C, YKL135C and YHL019C. However, its worth mentioning that two more proteins in this case will be wrongly predicted to be members of the complex namely YBR196C and YFR043C. To solve this problem, we propose to analyze the second rank neighbor. The two proteins in this case do not have more interactions with other members of the complex and should be ignored.

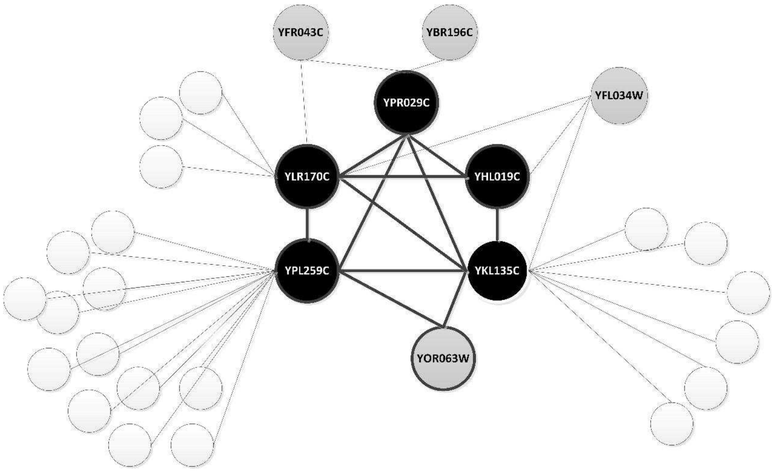

Figure 5: A type of a complex that was easily detected using ProRank.

In the future, we would like to study incorporating additional biological information of protein complexes. Sequence similarity has contributed to the accuracy of ProRank however, the improvement is not significant. The analysis on Cmplx-D1 shows that 79% of the 81 complexes contain more than 50% proteins which share similarity percentage ≥ 60% however, 38% of the 990 proteins in PPI-D1 share high similarity with proteins which belong to other complexes in the network. Furthermore, we investigated the incorporation of Gene Ontology (GO) annotation however, no improvement is expected since only 61% of the 81 complexes contain more than 50% proteins from the same GO. To this end, a probabilistic calculation of the affinity score between two proteins [47] and incorporating co-immunoprecipitation data to identify sets of preys that significantly co-associate with the same set of baits [48] could further improve the performance of ProRank.

Relaying on the classical network node-and-edge representation, where proteins are nodes and interactions are edges shows only which proteins interact; not how they interact [49]. We also plan to use tools such as PRISM [49] to evaluate the PPI reliability based on structural and evolutionary similarity to known protein interfaces. Protein-protein interface structures could indicate which binding partners can interact simultaneously.

## Acknowledgment

The authors would like to acknowledge the assistance provided by the Research Support and Sponsored Projects Office and the Faculty of Information Technology at the United Arab Emirates University (UAEU). Special thanks to Piers Campbell and Maighdlin McKean for their valuable comments, Wu Min and Liu Guimei for providing the datasets, and Sawsan Khuri from the University of Miami (Florida, USA) for helpful discussion and information.

## References

[1] Vanunu O., Magger O., Ruppin E., Shlomi T., and Sharan R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol.*, 6(1):e1000641, 2010.

[2] Oti M., Snel B., Huynen MA., and Brunner HG. Predicting disease genes using protein-protein interactions. *J Med Genet*, 43:691–698, 2007.

[3] Acuner S.E., Keskin O., Nussinov R., and Gursoy A. Enriching the human apoptosis pathway by predicting the structures of protein-protein complexes. *J Struct Biol.*, page In Press, 2012.

[4] Tzakos A.G., Fokas D., Johannes C., Moussis V., Hatzimichael E., and Briasoulis E. Targeting oncogenic protein-protein interactions by diversity oriented synthesis and combinatorial chemistry approaches. *Molecules*, 16(6):4408–27, 2011.

[5] Chen K.C., Wang T.Y., and Chan C.H. Associations between hiv and human pathways revealed by protein-protein interactions and correlated gene expression profiles. *PLoS One*, 7(3):e34240, 2012.

[6] Lee J.M., Han J.J., Altwerger G., and Kohn E.C. Proteomics and biomarkers in clinical trials for drug development. *J. Proteomics*, 74(12):2632–41, 2011.

[7] Bader G. D. and Christopher W. H. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, 2003.

[8] Dongen S. *Graph clustering by flow simulation*. University of Utrecht, 2000.

[9] Andrew D. K., Przulj N., and Jurisica I. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020, 2004.

[10] Przulj N., Wigle D. A., and Jurisica I. Functional topology in a network of protein interactions. *Bioinformatics*, 20(3):340–348, 2004.

[11] Adamcsek B., Palla G., Farkas I. J., Derenyi I., and Vicsek T. CFinder: locating cliques and overlapping modules in biological networks. *J. Bioinformatics*, 22(8):1021–1023, 2006.

[12] Sylvain B. and Jacques H. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7(7):488, 2006.

[13] Guimei L., Wong L., and Chua H. N. Complex discovery from weighted PPI networks. *Bioinformatics*, 25(15):1891–1897, 2009.

[14] Leung H., XIANG Q., Yiu S. M., and Chin F. Predicting protein complexes from ppi data: A core-attachment approach. *Journal of Computational Biology*, 16(2):133–139, 2009.

[15] Nepusz T., Haiyuan Y., and Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods*, 9:471–472, 2012.

[16] Reguly T., Breitkreutz A., Boucher L., Breitkreutz B.J., and Hon G. Comprehensive curation and analysis of global interaction networks of saccharomyces cerevisiae. *J Biol.*, 5:11, 2006.

[17] Brin S. and Page L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks & ISDN Systems*, 30:107117, 1998.

[18] Bryan K. and Leise T. The $25,000,000,000 eigenvector: the linear algebra behind Google. *SIAM Review*, 48(3):569–581, 2006.

[19] Meyer C. D. Langville A. N. and Meyer C. *Matrix Analysis.* Princeton University Press, 2006.

[20] Ishii H. and Tempo R. A distributed randomized approach for the pagerank computation: Part 1. pages 3523–3528, 2008.

[21] Chen Y. and Xu D. Understanding protein dispensability through machine-learning analysis of high-throughput data. *Bioinformatics*, 21:575–581, 2005.

[22] Vallabhajosyula R.R., Chakravarti D., Lutfeali S., Ray A., and Raval A. Identifying hubs in protein interaction networks. *PLoS ONE*, 4(4):e5344, 2009.

[23] Miho H., Takashi I., and Kengo K. Identification of transient hub proteins and the possible structural basis for their multiple interactions. *Protein Sci.*, 17:72–78, 2008.

[24] Kuang R., Weston J., Noble W. S., and Leslie C. S. Motif-based protein ranking by network propagation. *Bioinformatics*, 21(19):3711–3718, 2005.

[25] Brun C. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol.*, 5:R6, 2003.

[26] Hon N. C., Sung W. K., and Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630, 2006.

[27] Chua H. N., Ning K., Sung W. K., Leong H. W., and Wong L. Using indirect protein-protein interactions for protein complex prediction. *J. Bioinformatics and Computational Biology*, 6(3):435–466, 2008.

[28] Zaki N. M., Lazarova-Molnar S., El-Hajj W., and Campbell P. Protein-protein interaction based on pairwise similarity. *BMC Bioinformatics*, 10:150, 2009.

[29] Zaki N. M., Wolfsheimer S., Nuel G., and Khuri S. Conotoxin protein classification using free scores of words and support vector machines. *BMC Bioinformatics*, 12:217, 2011.

[30] Saigo H., Vert j., Ueda N., and Akutsu T. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689, 2004.

[31] Pearson W. R. and Lipman D. J. Improved tools for biological sequence comparison. In *Proc. Natl. Acad. Sci.*, volume 85, pages 24444–24448, 1988.

[32] Smith T. F. and Waterman M. S. Identification of common molecular subsequences. *Journal of Molecular Biology*, Vol. 147:195–197, 1981.

[33] Horn R. A. and Johnson C. R. *Matrix Analysis.* Cambridge Univ. Press, 1985.

[34] Zhang N. and Bilsland E. Contributions of saccharomyces cerevisiae to understanding mammalian gene function and therapy. *Methods Mol Biol.*, 759:501–23, 2011.

[35] Gavin A. C. et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636, 2006.

[36] Ho Y. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415:180–183, 2002.

[37] Gavin A. C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.

[38] Krogan N.J. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440:637–643, 2006.

[39] Uetz P. et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403:623–627, 1999.

[40] Ito T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. of The National Academy of Sciences*, 98:4569–4574, 2001.

[41] Mewes H. W. et al. Mips: a database for genomes and protein sequences. *Nucleic Acids Res.*, 28:37–40, 2000.

[42] Mewes et al. MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Research*, 34(Database-Issue):169–172, 2006.

[43] Brohee S. and van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7:488, 2006.

[44] Xionglei H. and Jianzhi Z. Why do hubs tend to be essential in protein network. *PLoS Genetics*, 2(6):1371, 2006.

[45] Warren G and Malhotra V. The organisation of the golgi apparatus. *Curr. Opin. Cell Biol.*, 10:493–498, 1998.

[46] Howe ML. et al. Transcription factor iia tau is associated with undifferentiated cells and its gene expression is repressed in primary neurons at the chromatin level in vivo.. *Stem Cells Dev.*, 2:175–90, 2006.

[47] Xie Z., Kwoh C. K., Li X. L., and Wu M. Construction of co-complex score matrix for protein complex prediction from ap-ms data. *Bioinformatics*, 27:i159i166, 2011.

[48] Geva G. and Sharan R. Identification of protein complexes from co-immunoprecipitation data. *Bioinformatics*, 27:111117, 2011.

[49] Kuzu G., Keskin O., Gursoy A., and Nussinov R. Constructing structural networks of signaling pathways on the proteome scale. *Curr Opin Struct Biol.*, 22:1–11, 2012.