# Protein Protein Interaction Identification using Weighted Graphs by Random Walk

M.Thillainayaki, M.Hemalatha

**Abstract**— Proteins in the same protein complexes should highly interact with each other but rarely interact with the other proteins in protein-protein interaction (PPI) networks. All interaction network weighting schemes have been proposed so far in the literature in order to eliminate the noise inherent in interactome data. Visualization representation of data visually and is an important task in scientific research. PPI are discovered using mass spectrometry, or in silico predictions tools, resulting in large collections of interactions stored in specialized databases. Using Random walk on weighted graphs for identifying the interaction easily between Protein subsets and measuring the evaluation performance of proteins, Graphs for PINs visualizing the high number of nodes and connections, the heterogeneity of nodes (proteins) and edges (interactions), the possibility to annotate proteins and interactions with biological information that enriches the PINs with semantic information, and maintained as a separate databases for easy retrieval information of proteins from various Protein databases.

**Index Terms**— Graph, Networks, Protein Protein interaction, Random Walk, Weighted Networks

## 1 Introduction

Protein interactions can also be classified into two types based on their timing and the spatial distribution of binding sites on the protein surface. Much effort has been devoted to propose computational approaches for detecting PPIs based on various data types, such as genomic information, protein domain and protein structure information. For example, Yu et al. proposed a method based on secondary structures for inferring PPIs, and found that helix and disordered structures account for most of interacting regions. Products of coexpressed genes may form stable complexes and interact with each other simultaneously, which is only possible when a network hub ("party hub") possesses a unique binding site for each interaction partner. Alternatively, hub proteins that are not co-expressed with their interaction partners are believed to bind their partners individually at different times (or in different cellular locations) via the same interface ("data hubs"). Following Kim et al.[4] we refer to the interactions of the first and the second type as simultaneously possible (SP) and mutually exclusive (ME), respectively. SP and ME interactions and the corresponding binding interfaces can be directly studied by overlaying highquality protein interaction data with known threedimensional structures of protein complexes. Analyses of such a structurally resolved interaction network (SIN) together with gene expression patterns revealed distinctly different cellular roles of party and date hubs, with the former corresponding to stable network modules and the latter connecting modules with each other. Date hubs show much lower average degree and are more often encoded by essential genes than party hubs. It is of great significance to develop computational methods by only using protein sequence information for predicting protein-protein interactions. Current computational systems for predicting PPIs usually consist of two parts, feature extraction and machine learning model.

However, it is often hard to distinguish between these two structures by relying only on PPIN, as in general the analyzed protein interactions do not have temporal and spatial information. Nevertheless, since PPIN represent undirected binary or weighted graphs, several graph-based inference approaches have been successfully employed to detect modularity. The majority of such approaches evaluate interactome topological features, and typical examples are node degree and clustering coefficient, both based on the levels of connectivity of each node. Both global and local connectivity can be explored by these methods, depending on the kind interactome analysis to be performed. The results may vary, as methods are based on different principles. For instance, the two main contributions to our work come from the application of two algorithms, CFinder and MCODE. Interestingly, they deal with network modularity through similar topological instruments, but achieve quite different outcomes; therefore, we based our analysis on them, while also evaluating other methods. In parallel, a substantial heterogeneity of human interactome datasets has been generated depending on the underlying methods of identifying and characterizing protein interactions. Besides high-throughput approaches, in particular the curation of literature and the provision of computational predictions have allowed for the mapping of the human interactome.

## 2.1 Protein interaction

Main requirements for the visualization of protein interaction networks are: Clear rendering of network structure and substructures, such as dense regions or linear chains; Fast rendering of huge networks; Easy network querying through focus and zoom; Compatibility with the heterogeneous data formats used for PIN representation; Interoperability with PPI databases, allowing the automatic querying of single or multiple databases using existing middlewares (e.g. cPath [6]); Integration of heterogeneous data sources, e.g. functional information about proteins extracted from biological ontologies. Most studies have assumed that proteins in a protein complex form a highly connected subnetwork in the PPI network. However, although proteins in the same complex are more highly connected with each other than with proteins from different complexes, many protein complexes are not very dense. A statistical analysis was used to verify this point. The density of a protein complex can be defined as

$$\text{Density} = \frac{2|E|}{|V|(|V|-1)} \qquad 1$$

where |V|j is the number of proteins in the protein complex and |E| is the number of protein-protein interactions in the protein complex. The maximum density is 1 and the minimum density is 0. The analysis used an unweighted protein interaction network of yeast downloaded from the DIP database, which contained 5093 proteins and24 743 interactions [1]. The 408 known yeast protein complexes from the MIPS database were used as benchmark complexes with 172 protein complexes with 2 proteins and 236 protein complexes with more than 2 proteins. The statistical analysis considered the complexes with more than 2 proteins. The results are shown in Fig. 1.

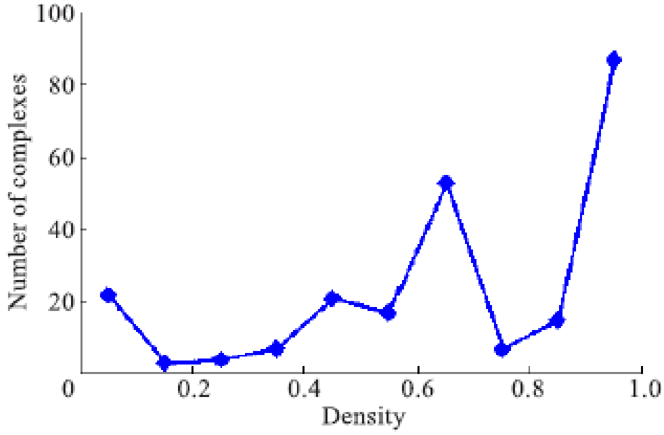


**Figure 1. Number of complexes with different densities**

To mine protein complexes from protein-protein interaction networks, first proteins in one complex should be in a same subcellular location. Thus, proteins are clustered if they are in a same subcellular location. If some proteins do not have a subcellular location annotation, these proteins may appear in any subcellular location. The subcellular locations of the yeast protein data was downloaded from the Biocomp database. Second, a topology property was defined to decide whether two proteins were in the same complex. This property was defined as

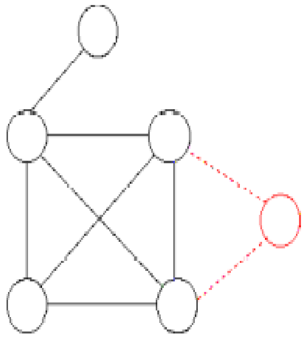
$$np = \sqrt{\frac{|n(u) \cap n(v)^2|}{|n(u)||n(v)|}} \qquad 2$$

where np represents the percentage of same neighbors of two proteins, n(u) is the neighbor set of protein u, and n(v) is the neighbor set of protein v. The rule gives manypairs of proteins where each pair of proteins is regarded as a cluster. Then these proteins are used to construct a graph with the following two steps to refine clusters.

| | Organism | | | | | |
|---|---|---|---|---|---|---|
| | *Homo Sapiens* | | *Mus Musculus* | | *Saccharomyces C* | |
| **Databases** | *# Nodes* | *# Edges* | *# Nodes* | *# Edges* | *# Nodes* | *# Edges* |
| *I2D* | 14847 | 156188 | 12818 | 145119 | 11194 | 152877 |
| *INTACT* | 18161 | 86537 | 8305 | 18896 | 8958 | 105440 |
| *MINT* | 8624 | 26698 | 8624 | 26698 | 62621 | 5661 |
| *DIP* | 3337 | 4794 | 1361 | 1468 | 5087 | 24114 |
| *BIOGRID* | 15239 | 125045 | 5142 | 11565 | 6248 | 304198 |

**Table 1 : Dimensions of PIN in some organisms.**

**Step 1** If a connected component has more than three proteins and if one protein's degree is 2 and it can form a triangle with other proteins, the two clusters represented by outside edges (dotted line) are deleted with the proteins on the triangle add as one seed.



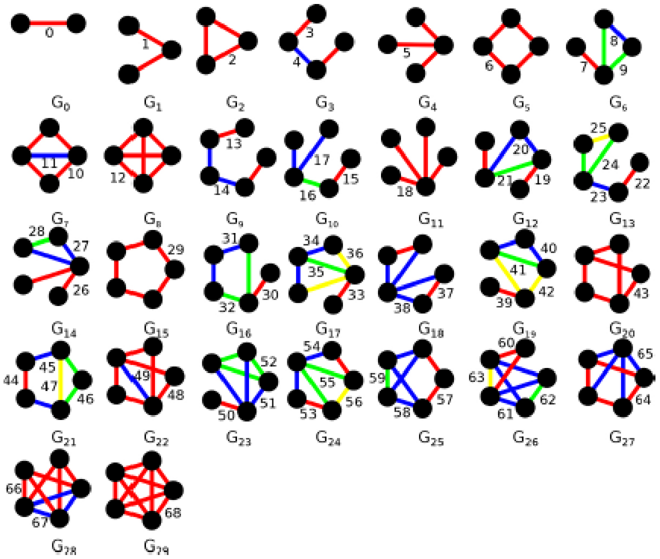


Figure. 2 Connected components.

**Step 2** If a connected component forms a clique as in Fig. 2, then all the seeds represented by the edges are deleted and all the proteins in the clique are combined into one seed.

## 2.2 Random Walk Steps

A novel random walk method for protein subcellular localization based on amino acid composition. By mapping the protein data into a weighted and partially labeled graph where each node represents a protein sequence, a random walk classification model to predict labels of unlabeled nodes based on a theoretical model. Random walk classifier was coded in MATLAB. Given the training data and their classes, compute the state matrix Y and weight matrix W. The similarity or weight between two nodes was given according to the radius basis function. First the complete graph offers each labela chance to reach the unlabeled node in at atleast one step. Second, the good accuracy gradually declines after the peak value of t. Since the labeld training data is often deterministic, the transistion matrix built over the labeled data is commonly treated as a unit matrix in semisupervised random walk methods.

We next aimed to deduce a simple classifier based on the nodes that are labeled so it can be applied to predict the labels of the unlabeled nodes. Our solution was a state vector y that provides the label for an unlabeled data point x. We first provide an example to clarify the process of label propagation through random walks. Consider an initial graph G constructed over the training data (X, Y) = {(x1, c1), (x2, c1), (x3, c2)}. Each data point lacking a label is added into graph G as an unlabeled node. Figure 3 displays such a graph G′ after three unlabeled data points were added. The graph G′ is often assumed to be label-connected to become completely labeled that is, it is possible to reach a labeled node from any unlabeled node in a finite number of steps. For example, if in a random walk, the sixth node v6 ends at the second node v2, then this node will be labeled as c1. Node classification relies on a random walk originating at the unlabeled node vj and ends at one labeled node vi after several steps, and in this way, vj obtains its label from vi. If during the walk an unlabeled node reaches a labeled node for the first time, it will not remain at that node because the labeled nodes are not absorbing states; rather, the unlabeled node will move to another node with a certain probability. Since graphs G and G′ are undirected and symmetric, a random walk that starts at vj and ends at vi can be also revertible. Next, we assume $p(v_i, v)$ to be the state-transition probability with which a walk proceeds from node vi in V to the new node v represented by unlabeled data point x. The state y of new node v is represented as

$$\gamma = \sum_{v_i \in V} p(v_i, v)\gamma_i = Yp_v \qquad 3$$

where

$$p_v \underline{\underline{def}} = p(V,v) = \begin{bmatrix} p(v_1,v) \\ p(v_2,v) \\ . \\ . \\ . \\ p(v_n,v) \end{bmatrix} \qquad 4$$

The idea underlying the random walk methods is that the probability of labeling a node v with a label (or state) y is the total probability that a random walk starting at v will end at a node labeled y. F(x) therefore is more likely to return a probability distribution such as $F(x_i) = F(v_i) = [f_{1i}, f_{2i}, ..., f_{ci}]T$, where each distribution $f_{ji}$ refers to the total probability that the a random walk starting at node vistops at any node labeled $c_j$ after t steps. The largest $f_{ji}$ allows $v_i$ to be assigned label $c_j$.

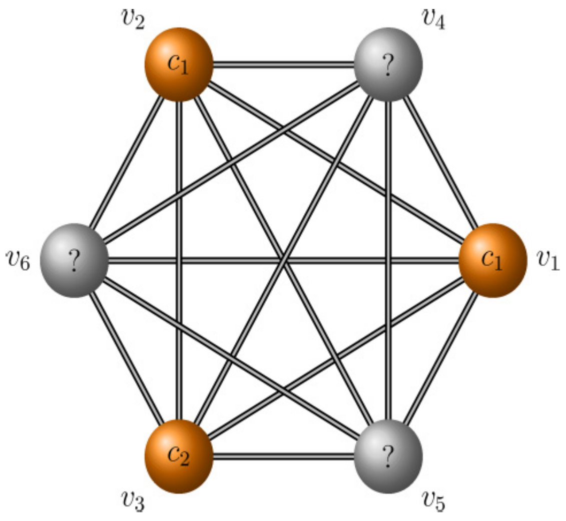


**Figure 3. Partially labeled graph**

## 2.3 Evaluation Measures

To evaluate the performance of the proposed method, the following criterion was used: the overall prediction accuracy, sensitivity, precision, and correlation coefficient was calculated.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad 5$$

$$SENS = \frac{TP}{TP + FN} \quad 6$$

$$PREC = \frac{TP}{TP + FP} \quad 7$$

$$CORREFF = \frac{TPxTN - FPxFN}{\sqrt{(TP + FN)x(TN + FP)x(TP + FP)x(TN + FN)}} \quad 8$$

where true positive (TP) denotes the number of true samples which are predicted correctly; false negative (FN) is the number of samples predicted to be noninteracting pairs incorrectly; false positive (FP) is the number of true non-interacting pairs predicted to be PPIs falsely, and true negative (TN) is the number of true non-interacting pairs predicted correctly. Furthermore, the ROC curve was also calculated to evaluate the performance of proposed method. Summarizing ROC curve in a numerical way, the area under an ROC curve (AUC) was computed. Global encoding (GE) of protein sequences could be obtained by the following steps.

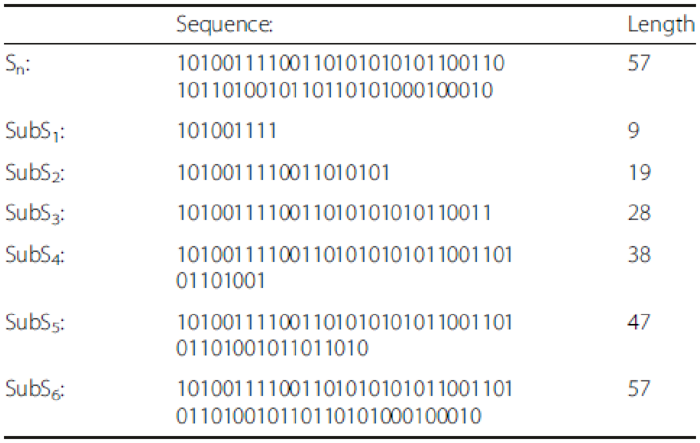
**Step 1.** Transformation of protein sequence Researches [2,3] have pointed out that amino acids can be classified into 6 difrent classes according to the physicochemical characteristic such as residues' hydrophobic property, charged property and so on. For the reduction of data complexity, we first encode the protein sequence substituting every amino acid by its class accordingly, and the substitution rules are presented in Table 2.

| Amino acid classification | |
|---|---|
| Aliphatic amino acid | C1 = {A,V,L,I,M,C} |
| Aromatic amino acid | C2 = {FW,Y} |
| Polar amino acid | C3 = {S,TN,Q} |
| Positive amino acid | C4 = {K,R} |
| Negative amino acid: | C5 = {D,E} |
| Special conformations | C6 = {G,P} |

**Table 2: Amino acid classification**

In this way, every protein sequence is represented by six symbols: C1, C2…C6. Based on this classification, we can further divide these 6 classes into 2 subsets each of which contains 3 different classes. By doing this, ten modes can be obtained as follows: {C1, C2, C3} vs {C4, C5, C6}, {C1, C2, C4} vs {C3, C5, C6}, {C1, C2, C5} vs {C3, C4, C6}, {C1, C2, C6} vs {C3, C4, C5}, {C1, C3, C4} vs {C2, C5, C6}, {C1, C3, C5} vs {C2, C4,C6}, {C1, C3, C6} vs {C2, C4, C5}, {C1, C4, C5} vs {C2, C3, C6}, {C1, C4, C6} vs {C2, C3, C5} and {C1, C5, C6} vs {C2, C3, C4}. We then transform every protein sequence into ten binary sequences based on these ten modes correspondingly.

**Step 2.** Partition of characteristic sequences In this step, every characteristic sequences are further divided into subsequences of different lengths by a special strategy. For any characteristic sequence $S_n = s_1, s_2,\ldots,s_n$ of length n, given a positive integer, Sn will be divided into L subsequences. We call the kth subsequence as SubSk (k = 1, 2,…, L) and SubSk is composed of the first ⌊kn/L⌋ numbers of Sn. Here we present an example to explain the process of characteristic sequence partition in Table 3.



| | Sequence: | Length |
|---|---|---|
| $S_n$: | 10100111100110101010101100110<br>1011010010110110101000100010 | 57 |
| $SubS_1$: | 101001111 | 9 |
| $SubS_2$: | 1010011110011010101 | 19 |
| $SubS_3$: | 1010011110011010101010110011 | 28 |
| $SubS_4$: | 101001111001101010101011001101<br>01101001 | 38 |
| $SubS_5$: | 101001111001101010101011001101<br>01101001011011010 | 47 |
| $SubS_6$: | 101001111001101010101011001101<br>011010010110110101000100010 | 57 |

**Table 3: Characterisitic sequence partition**

In this sample, the length of the given sequence is 57 and parameter L is set to be 6. So the length of its subsequences is 9, 19, 28, 39, 47 and 57 respectively.

**Step 3**. Extraction of feature vectors In the last step, feature vectors of composition and transition descriptors will be extracted from the subsequences produced in the prior step. The composition descriptor describes the frequencies of '0' and '1' in each subsequence. As a composition descriptor of one subsequence contains two frequency values, any characteristic sequence would be represented by a 2*L dimensional feature vector by the composition descriptor. Transition, as the second descriptor, account for the switch frequency between '0' and '1' in every subsequence. The times where '0' follows 1' and '1' follows '0' happen are counted independently.

It shows the process of descriptors' extraction from the subsequence 3 in the Table 2. The length of example sequence is 28; the numbers of '0' and '1'are 12 and 16 respectively; the transition times of '1-0' and '0-1' are both 9. Therefore, two values of composition descriptor are 12/28 = 42.86 % and 16/28 = 57.14 % respectively. The value of transition descriptor is 9 + 9 = 18. In this work, L is set to be 5 after adjusting for the best performance. As a protein sequence would be first transformed into 10 numerical sequences and each sequence would further be partitioned by 5 subsequences which can be represented by 3-dimension feature descriptors, the length of the whole feature vector of a protein sequence is 10*5*3 = 150.

It is often hard to distinguish between these two structures by relying only on PPIN, as in general the analyzed protein interactions do not have temporal and spatial information. Nevertheless, since PPIN represent undirected binary or weighted graphs, several graph-based inference approaches have been successfully employed to detect modularity. The majority of such approaches evaluate interactome topological features, and typical examples are node degree and clustering coefficient, both based on the levels of connectivity of each node.

**Algorithm:**

1. **Input**: Samples matrix and any test sample as query and target network, pairwise node and similarity score.
2. Normalize the columns of x to have a unit as a data preprocessing such as removal of non homologous nodes and inserting primary edges.
3. Calculate the set of nearest nodes, distance and name the solved node as in the subset.
4. Solve the table.
5. Compute each residual.
6. **Output:** Identity will be occurred for matching best in the target network for the given query.

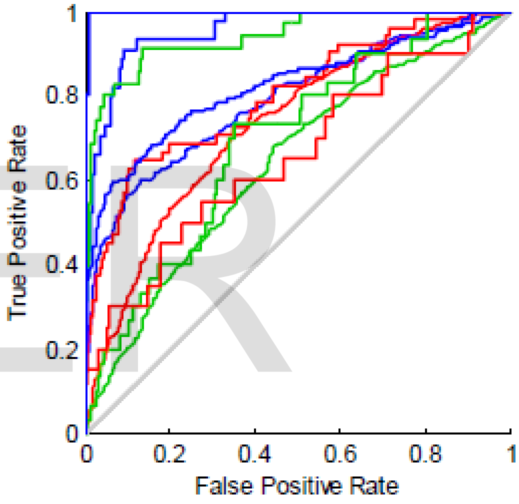


**Figure 4. ROC on Random Walk in different protein sequence.**

To solve the PPI inference problem, a novel and fast optimization methodmusing linear programming to integrate multiple heterogeneous data from a protein databases. We note that the induced sparsity implies a poor identification power with regard to the resolution spectrum, especially for small and intermediate module sizes. Due to the retrieval of coarse resolution modules, whose large sizes depend on incremental merging of small modules, a weakness of the MaxMod approach concerns its possible interpretation in biological applications. In addition, more reasons of concern exist with reference to methodological aspects. First, more than one partition could reach the maximal modularity (local maxima). Second, the modularity definition could reveal only some groups (due to bias). Third, as modularity calculation is sensitive to noise, an optimal partition may not be achieved. Consequently, the

MaxMod suboptimality effect of limiting the coverage for the network resolution spectrum requires investigation when all module sizes may in principle count.

## 3 Conclusion

A method named SiS (Significant Subnetworks) that finds the most probable subgraphs in a large biological network data set. SiS initializes a weighted graph named the template graph that summarizes the input graphs. SiS takes advantage of the template graph while finding the most probable subgraphs of a user-given size, k. In other words, SiS finds the subgraphs of k interactions with the largest probability to appear in a network selected randomly from the input data set. Multiple heterogeneous features for the proteins in the network are then combined into the form of weighted kernel fusion, which provides a new "adjacency matrix" for the whole network that may consist of disconnected components but is required to comply with the transition matrix on the training subnetwork. This requirement is met by adjusting the weights to minimize the element-wise difference between the transition matrix and the weighted kernels. The minimization problem is solved by linear programming. The weighted kernel fusion is then transformed to regularized Laplacian (RL) kernel to infer missing or new edges in the PPI network, which can potentially connect the previously disconnected components. An accurate 3D structure-independent computational method for classifying
Protein interactions into simultaneously possible (SP) and mutually exclusive (ME) as well as into obligate and non-obligate. Our classifier exploits features of the binding partners predicted from amino acid sequence, their functional similarity, and network topology. The method represents protein surface patches using labeled graphs and uses a graph kernel method to calculate the similarities between graphs. A new surface patch is predicted to be interface or non-interface patch based on its similarities to known DNA-binding patches and non-DNA-binding patches. The proposed method achieved high accuracy when tested on a representative set of 146 protein-DNA complexes using leave-one-out cross-validation. Then, the method was applied to identify DNA-binding sties on 13 unbound structures of DNA-binding proteins. In each of the unbound structure, the top 1 patch predicted by the proposed method precisely indicated the location of the DNAbinding site. Comparisons with other methods showed that the proposed method was competitive in predicting DNA-binding sites on unbound proteins. By modeling the optimization of the composite network and the prediction problems within a unified objective function. In particular, we use a kernel target alignment technique and the loss function of a network based classifier to jointly adjust the weights assigned to the individual networks. We show that the proposed method, called MNet, can achieve a performance that is superior (with respect to different evaluation criteria) to related techniques using the multiple networks of four example species (yeast, human, mouse, and fly) annotated with thousands (or hundreds) of GO terms,  novel multiple network alignment algorithm based on a context-sensitive random walk model. The random walker employed in the proposed algorithm switches between two different modes, namely, an individual walk on a single network and a simultaneous walk on two networks. The switching decision is made in a context-sensitive manner by examining the current neighborhood, which is effective for quantitatively estimating the degree of correspondence between nodes that belong to different networks, in a manner that sensibly integrates node similarity and topological similarity. The resulting node correspondence scores are then used to predict the maximum expected accuracy (MEA) alignment of the given networks. A novel statistical method to extract interacting residues and interacting patches can be clustered as predicted interface residues. In addition, structural neighboring property can be adopted to construct a new energy function, for evaluating docking solutions. It includes new statistical property as well as existing energy items. However, the novel measure based on profile-profile comparisons substantially improved the performance of the four methods, especially when very low sequence identity datasets were evaluated. We also performed a parameter optimization step to determine the best configuration for each clustering method. Random Walks to predict missing (or new) functions of partially annotated proteins. Particularly, we apply downward random walks with restart on the GO directed acyclic graph, along with the available functions of a protein, to estimate the probability of missing functions. To further boost the prediction accuracy. Comparing with nonessential proteins, essential proteins appear more frequently in certain subcellular locations and their evolution more conservative. By integrating the information of subcellular localization, orthologous proteins and PPI networks, we propose a novel essential protein prediction method.

## REFERENCES

[1] Xiaojun Ding Et al, “*Mining Protein Complexes from PPI Networks using the Minimum Vertex Cut*”, Tsinghua Science and Technology, Volume 17, Number 6, December 2012.

[2] Capra JA, Singh M. “*Predicting functionally important residues from sequence conservation*”. Bioinformatics. 2007;23(15):1875–82.

[3] Zhang ZH, Wang ZH, Wang YX. “*A new encoding scheme to improve the performance of protein structural class prediction*”, Advances in natural computation. Berlin: Springer; 2005. p. 1164–73.

[4] Kim PM, et al, “*Relating Data structures to protein networks provides evolutionary insights*” Science 2006, 314(5807), 1938 – 41.

[5] A.D.King, N. Przulj and I. Jurisica1, 2004, “*Protein complex prediction via cost-based clustering*”, Vol. 20 no. 17, pages 3013–3020.

[6] Agapito et al, 2013, “*BMC Bioinformatics Visualization of protein interaction networks: problems and solutions*”, 14(Suppl 1):S1.

[7] Arko Provo Mukherjee, Srikanta Tirthapura, 2016, ”*Enumerating Maximal Bicliques from a Large Graph using MapReduce Enumerating Maximal Bicliques from a Large Graph using MapReduce*”, DOI 10.1109/TSC.2523997.

[8] Bala zs Adamcsek, Gergely Palla, Illes J. Farkas, Imre Dere nyi and Tamas Vicsek, 2006, “*CFinder: locating cliques and overlapping modules in biological networks*”, Advance Access publication February 10.

[9] Cheng-Yu Yeh, Hsiang-Yuan Yeh, Carlos Roberto Arias, and Von-Wun Soo, 2012, “*Pathway Detection from Protein Interaction Networks and Gene Expression Data Using Color-Coding Methods and A∗ Search Algorithms*”, The ScientificWorld Journal Volume.

[10] Cody Hudson, Bernard Chen, and Dongsheng Che, 2014, “*Hierarchically Clustered HMM for Protein Sequence Motif Extraction with Variable Length*”, TSINGHUA SCIENCE AND TECHNOLOGY, pp 635-647 Volume 19, Number 6, December.

[11] Florian Baumann, Arne Ehlers, Karsten Vogt, Bodo Rosenhahn, 2012, “*Cascaded Random Forest for Fast Object Detection*”.

[12] Javad Zahiri, Joseph Hannon Bozorgmehr and Ali Masoudi-Nejad, 2013, “*Computational Prediction of Protein–Protein Interaction Networks: Algorithms and Resources*”, Current Genomics, 14, pp 397-414.

[13] Jia Hao Fan, jianer Chen and Sing Hoi Sze, 2012, “*Identifying Complexes from Protein Interaction Networks According to Different Types of Neighborhood Density*”, Journal of Computational Biology, Volume 19, Number 12.

[14] Jianxin Wang, Min Li, Huan Wang, and Yi Pan, 2012, “*Identification of Essential Proteins Based on Edge Clustering Coefficient*”, IEEE/ACM Transactions on Computational Biology and Bioinfomratics, Vol. 9, No. 4, July.

[15] Liang Zhao, Steven C.H. Hoi, Zhenhua Li, Limsoon Wong, Hung Nguyen, and Jinyan Li, 2012, “*Coupling Graphs, Efficient Algorithms and B-Cell Epitope Prediction*”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 11, No. 1, January/ February.

[16] Md. Altaf-Ul-Amin, Tetsuo Katsuragi, Tetsuo Sato, and Shigehiko Kanaya, 2015, “*Review Article - A Glimpse to Background and Characteristics of Major Molecular Biological Networks*”, Hindawi Publishing Corporation BioMed Research International Volume.

[17] Min Li, Jian-er Chen, Jian-xin Wang , Bin Hu and Gang Chen, 2008, “*Modifying the DPClus algorithm for identifying protein complexes based on new topological structures*”, BMC Bioinformatics.

[18] Peng et al., 2014, “*Improving protein function prediction using domain and protein complexes in PPI networks*”, BMC Systems Biology.

[19] Peng Liu,1 Lei Yang, Daming Shi, and Xianglong Tang, 2015, “*Prediction of Protein-Protein Interactions Related to Protein Complexes Based on Protein Interaction Networks*”, Hindawi Publishing Corporation BioMed Research International Volume.