Proceedings of the International Conference on Biomedical Engineering and Systems
Prague, Czech Republic, August 14-15, 2014
Paper No. 124

Multi-protein Complex Detection by Integrating Network Topological Features and Biological Process Information

Nazar Zaki
College of Information Technology, United Arab Emirates University (UAEU)
Al Ain 17551, UAE
nzaki@uaeu.ac.ae

Abstract - Predicting multi-protein complexes is becoming a central problem in system biology as it could be a way to reveal the biochemical functions of a protein. Most of the recently developed methods focus on topological information to detect multi-protein complexes. In this paper, biological and topological features characterizing protein complexes are extracted and used in conjunction with muti-class support vector machines to detect multi-protein complexes from the protein-protein interaction network. The proposed method was able to detect 76 complexes out of 81 reference complexes with high precision. In comparison with state-of-the-art methods, the evaluation results indicate that the applied method has great potential in detecting multi-protein complexes.

***Keywords*: ** Multi-class SVM, Protein complex, Network topological features, Cellular localization, Biological process.

1. Introduction

The use of high-throughput screening methods has contributed significantly to the growing amount of Protein-Protein Interaction (PPI) data. It is desirable to use this wealth of data to detect multi-protein complexes. A multi-protein complex is a group of two or more associated proteins formed by interactions that are stable over time. It was shown by Vanunu et al. (2010) that a protein and its high-confidence interactors are believed to form a putative multi-protein complex that is related to diseases such as prostate cancer, alzheimer's disease and type 2 diabetes. Therefore, several methods were developed to detect the multi-protein complexes.

Earlier methods include Markov clustering (MCL) (Dongen, 2000), restricted neighborhood search clustering (RNSC) (Andrew et al., 2004), CFinder [4] molecular complex detection (MCODE) algorithm (Bader and Christopher, 2003). Recent methods include Maximal Cliques (CMC) (Guimei et al., 2009) for discovering multi-protein complexes in weighted PPI networks. CMC uses an iterative scoring method called AdjstCD to assign weights to protein pairs. The AdjstCD weight in this method indicates the reliability of the interaction between protein pairs. Nepusz and Paccanaro developed ClusterONE (2012) which initiate from a single seed vertex before a greedy growth procedure begins to add or remove vertices in order to find clusters of proteins in the PPI with high cohesiveness. Zaki et al. developed ProRank (2012) which rank the importance of each protein in the network based on the interaction structure and the evolutionarily relationships between proteins. The ranking process in this case proved valuable for detecting multi-protein complexes. A recent method called PEWCC (Zaki et al., 2013) which based on the concept of weighted clustering coefficient has shown great potential in detecting multi-protein complexes.

Most of the above mentioned methods mainly focus on topological information and fail to consider the information from protein primary sequence which is of considerable importance for multi-protein complex detection. It was mentioned by the authors of ProRank (Zaki et al., 2012) that the accuracy improvement achieved by incorporating sequence similarity information to their algorithm is not significant. Therefore, to achieve a breakthrough, we need a deeper understanding of the characteristics of the proteins within these complexes. In this paper, we propose a supervising learning method for multi-protein complex detection by integrating network topological features and biological process information to be used in conjunction with multi-class support vector machines.

2. Method

Our method which we call it SVM-Net mainly consists of four major steps, feature extraction, preparing the data (pre-processing), mining patterns (classification) and post-processing. These steps are described in the proceeding sections and are illustrated in Figure 1.

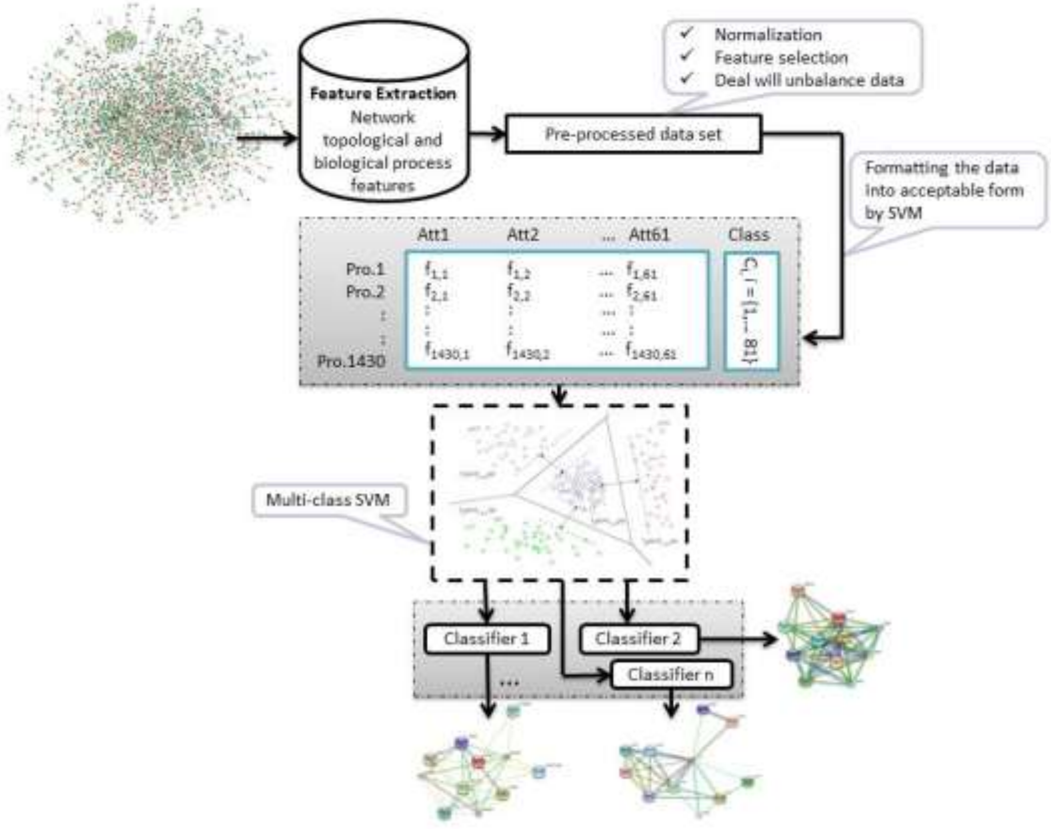
Feature Extraction
Network topological and biological process features
Pre-processed data set
✓ Normalization
✓ Feature selection
✓ Deal will unbalance data
Formatting the data into acceptable form by SVM
Att1
Att2
... Att61
Class
Pro.1
Pro.2
Pro.1430
$f_{1,1}$ $f_{1,2}$ ... $f_{1,61}$
$f_{2,1}$ $f_{2,2}$ ... $f_{2,61}$
$f_{1430,1}$ $f_{1430,2}$ ... $f_{1430,61}$
$C_i \mid = \{1,\ldots 81\}$
Multi-class SVM
Classifier 1
Classifier 2
Classifier n
...

Fig. 1. Overview of the proposed SVM-Net method.

2. 1. Features Extraction

The formation of multi-protein complexes might be regulated at different levels, including transcriptional regulation. In prokaryotes for instance a significant proportion of the genes that are co-regulated at the transcriptional level usually code for proteins that physically interact (Simonis et al., 2004). This proportion is even higher for gene groups whose co-regulation is conserved in different genomes (Huynen et al., 2000). Therefore, two sets of valuable features can be extracted. The first is the out-degree and in-degree related to transcriptional regulation interaction. This feature represents the number of outgoing or incoming links to the gene g corresponding to a protein. Links in this case are represented in terms of transcriptional regulation interactions. The second is the betweenness centrality with respect to transcriptional regulation interactions. If $N_{g_1 g_2}$ denotes the number of shortest paths between two genes g_1 and g_2 then the value $N_{g_1 g_2}(g)$ can be defined as the number of shortest paths between g_1 and g_2 passing through g. The paths in this case are represented in terms of transcriptional regulation interactions. Similarly, betweenness centrality with respect to the physical interaction can also be calculated.

The graph abstraction of protein interactions is crucial for the understanding of the global behaviour of the network and therefore integrating topological properties, cellular components, and biological processes retains valuable knowledge of the characteristics of the multi-protein complexes. Hence, two informative set of features can be extracted which include the cellular components (cytoplasm, endoplasmic reticulum, mitochondrion, nucleus or other localization) and biological processes (cell cycle, metabolic process, signal transduction, transcription, transport or other process). The above four feature sets were obtained from Acencio and Lemke (2009).

PPI is often represented as a graph $\mathrm{G} = (\mathrm{V}, \mathrm{E})$, where V is a set of nodes (proteins) and E is a set of edges (interactions) connecting pairs of nodes. This representation allows us to study the network using the concepts and principles of graph theory. Therefore, two sets of features such as betweenness centrality related to integrated functional, degree related to integrated functional, maximum neighborhood component and density of maximum neighborhood component are considered. In case of the betweenness centrality and the degree, the values are represented in terms of integrated functional with respect to physical interaction (PI) and genomic context (GC) network interactions. Following Chiou-Yi Hor et al. (2013), the network information are collected from Hu et al. (2009) and the features are calculated using iGraph software (Csárdi and Nepusz, 2006). Maximum neighborhood component (MNC) and density of maximum neighborhood component (DMNC) properties were proposed by Lin et al. (2008) and Chin (2010). Other topological feature such as Clique level was also calculated. The clique level (Hwang et al., 2009) of protein i is defined as the maximal clique containing i. Here, only cliques with sizes between 3 and 10 proteins were taken into consideration.

Sequence primary structural features such as protein length, cysteine count, amino acid occurrence, average cysteine position, average distance of every two cysteines, cysteine odd-even index, average hydrophobicity, average hydrophobicity around cysteine, cysteine position distribution and average PSSM of amino acid were also used. All the above 10 protein features were taken from Lin et al. (2010) and were used to detect essential proteins from PPI (Chiou-Yi et al., 2013).

Evolutionary related feature namely the phyletic retention was also considered. In this case the phyletic retention of protein i is the number of organisms in which an ortholog is present. The ortholog of each protein was obtained from Hwang et al. (2009). The Number of paralagous genes which defined as the number of genes that are present in the same genome and the open reading frame (ORF) length were also considered.

2. 2. Preparing and Pre-processing of the Data

The data extracted are often very sparse and therefore, standardization of the dataset is a common requirement for many machine learning estimators as they might perform badly if an individual feature is not normally distributed. In this case all features should be normalized and scaled between 0 and 1.

To insure that all our attributes are meaningful, a feature selection is used to assess the relevance of each attribute. We focus on using a feature selection method based on filtering. Filtering algorithms use independent search and evaluation method to determine the relevance of features variables to the data mining task. Therefore we employed the “GainRatioAttributeEval” method available in Weka (Mark et al., 2009) to evaluate the worth of an attribute by measuring the gain ratio with respect to the class. The distribution of proteins across complexes is obviously imbalance and therefore, a resampling function is used.

2. 3. Mining Patterns (Classification)

Once the data is pre-processed a sensible data mining task must be designed to comply with the objectives of predicting proteins in the multi-protein complexes. This problem can be handled by utilizing a multi-classification technique and therefore, Support Vector Machines (SVM) (Vapnik, 1998; Cristianini and Shawe-Taylor, 2000) was selected to be used. The basic idea of the SVM algorithm is to map the given training set into a possibly high-dimensional feature space and attempting to locate in that space a hyperplane that maximizes the distance separating the positive from the rest of the examples.

The SVM algorithm addresses the general problem of machine learning to discriminate between positive and negative examples of a given class of n-dimensional vectors. In order to discriminate between proteins across complexes, the SVM learns a classification function from a set of positive examples μ+ and set of negative examples μ-. The classification function takes the form:

$$f(x) = \sum_{i:x_i \in \mu+} \lambda_i K(x, x_i) - \sum_{i:x_i \in \mu-} \lambda_i K(x, x_i) \qquad (1)$$

where the non-negative weights λ_i are computed during training by maximizing a quadratic objective function and the function $K(.,.)$ is called a kernel function. Any accident case x is then predicted to be positive if the function $f(x)$ is positive. More details about how the weights λ_i are computed and the theory of SVM can be found in (Vapnik, 1998; Cristianini and Shawe-Taylor, 2000). In the case of multiclass SVM labels which are drawn from a finite set of several elements are assigned to the instances. The dominant approach for doing so is to reduce the single multiclass problem into multiple binary classification problems.

2. 4. Post-processing Patterns

Following the classification step it is important to evaluate the patterns detected by the SVM. Several evaluation measures are used in this study such as Precision ($\frac{\text{TP}}{\text{TP+FP}}$), Recall ($\frac{\text{TP}}{\text{TP+FN}}$), F1 measure ($2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision+Recall}}$) and Accuracy ($\frac{\text{TP+TN}}{\text{All}}$), where TP, TN, FP, FN and All are defined as:

TP: related protein classified as "related".
TN: unrelated protein classified as "related".
FP: related protein classified as "unrelated".
FN: unrelated protein classified as "unrelated".
All: total number of proteins in the dataset.

Once all proteins are classified in groups (complexes) with reasonable classification accuracy it is important to assess the quality of the detected complexes. To evaluate the accuracy of the detected complexes, we used the Jaccard index which is defined as follows:

$$Match(K, R) = \frac{|V_K \cap V_R|}{|V_K \cup V_R|} \qquad (2)$$

where V_K and V_R are the set of proteins in complex K and R, respectively. The complex K is defined to match the complex R if $Match(K, R) \geq \alpha$ where $\alpha = \{0.3 \; or \; 0.5\}$ (as most of the available methods were evaluated).

To estimate the cumulative quality of the detection, we follow Zaki et al. [8] and compare the number of matching complexes with the number of reference complexes using recall ($RE_c = \frac{N_{MK}}{N_K}$), precision ($PR_c = \frac{N_{MR}}{N_R}$) and F_c-measure ($F_c = 2 \times \frac{PR_c \times RE_c}{PR_c + RE_c}$), where N_{MK} is a number of matching reference complexes, N_{MR} is a number of detected reference complexes, N_K is a number of reference complexes and N_R is the number of detected complexes.

To assess the accuracy estimation of the proteins predicted in the reference and detected complexes three further characteristics are used:

$$\text{Recall: } RE_N = \frac{\sum_{i=1}^{N_{MK}} |C_i|}{\sum_{i=1}^{N_K} |K_i|}, \text{ where } |C_i| = max_{R_j:Match(K_i,P_j) \geq \propto} |K_i \cap R_j| \qquad (3)$$

$$\text{Precision: } PR_N = \frac{\sum_{i=1}^{N_{MR}} |C_i|}{\sum_{i=1}^{N_R} |R_i|}, \text{ where } |C_i| = max_{K_j:Match(K_i,P_j) \geq \propto} |R_i \cap K_j| \qquad (4)$$

F_N-measure: $(F_N = 2 \times \frac{PR_N \times RE_N}{PR_N + RE_N})$ (5)

Calculations were made of precision and recall at complex and complex protein levels. Furthermore, we evaluated the performance of our method using the maximum matching ratio (MMR) which reflects the maximal one-to-one mapping between detected and reference complexes. The algorithm to calculate the MMR is available from (Nepusz and Paccanaro, 2012).

3. Experimental Work and Results

The effectiveness of the proposed method is evaluated using a PPI dataset which was prepared by Gavin et al. (2006). The dataset contains 1430 proteins, 6531 interactions, with network density of 0.006, and average number of interactions equal to 9.134. The network contains no isolated nodes and a diameter of 13. The reference data of complexes was created from MIPS (Mewes et al., 2002). In the case of MIPS, only complexes that were manually annotated from DIP interaction data are considered. Following Leung et al. (2009), complexes of sizes less than 5 proteins are excluded and therefore, 81 complexes were considered.

The experimental work started by the exploration and the preparation of the PPI dataset. A total of 61 features were extracted (as explained in section 2). The features were analysed and the “GainRatioAttributeEval” method reveals that features related to organelle such as vacuole, mitochondrion and endoplasmic reticulum are strongly linked with the detection of multi-protein complexes. Features related to amino acid occurrence and in particular “GLN”, “GLY” and “LYS” are also proved valuable. Similarly, there is no evidence suggesting that organelle such as peroxisomes and the bud neck (a constriction between the mother and the daughter cell (bud) in an organism that reproduces by budding) have no strong links to the characterization of multi-protein complexes.

One other observation inferred from the data exploration as shown in Figure 2 that the distribution of proteins across multi-protein complexes is unbalanced. From data mining point of view this data requires balancing and therefore, resampling method with random seed equal to 1 was used. The resampling in this case produces a random subsample of a dataset using replacement.

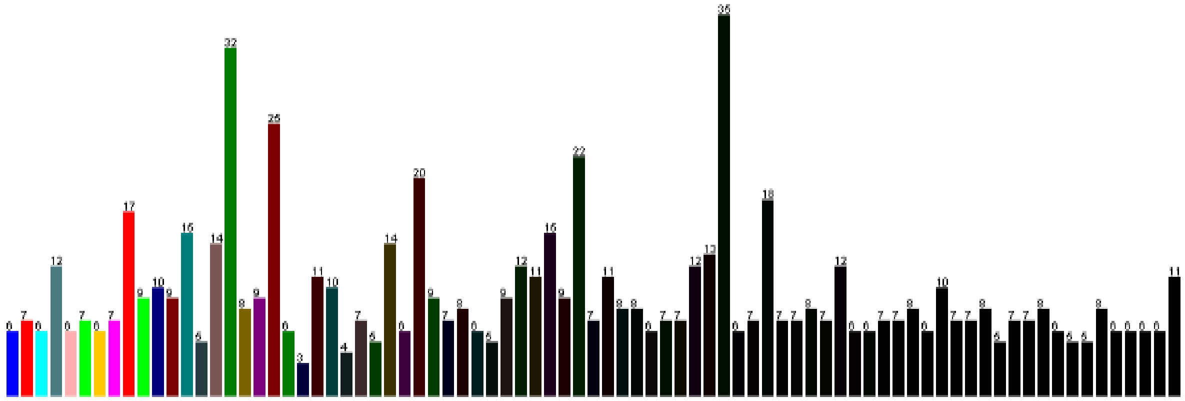

Fig. 2. The distribution of proteins across complexes.

Once the pre-processing step is completed and the dataset is prepared, multiclass SVM was used. The Library for Support Vector Machines (LibSVM) implemented by Rong-En and Chih-Jen Lin (2005) was used. One of the significant parameters needed to tune the SVM is the choice of the kernel function. The kernel function allows SVM to locate the hyperplane in high dimensional space that effectively separate the training data (Zaki et al. 2011). The Gaussian Radial Basis function (RBF) was used as it allows pockets of data to be classified which is more powerful way than just using a linear dot product.

To know how accurately our predictive model will perform in practice, 10-fold cross validation was used. The overall classification accuracy of assigning proteins to their corresponding complex is 71.05. The classification Precision, Recall and F1 measure are 0.72, 0.71 and 0.71, respectively. The list of the detected complexes was then compared to the reference complex dataset (both available at http://faculty.uaeu.ac.ae/nzaki/Research.htm). The proposed method was able to impressively detect 76 complexes out of the 81 reference complexes with the value of $\alpha = 0.30$. Furthermore, we compared the performance of SVM-Net to other state-of-the-art methods for detecting multi-protein complexes. The comparison is shown in Table 1. More than one quality measures were used to assess the performance of each algorithm. The parameters of the methods listed in Table 1 were optimized to achieve the best accuracy possible.

Table 1. Performance comparison of SVM-Net to ClusterONE, CMC, MCode, PEWCC and ProRank.

Evaluation Measures	ClusterONE	CMC	MCode	PEWCC	ProRank	SVM-Net
RE$_c$	0.803	0.753	0.568	0.753	0.79	**0.938**
PR$_c$	0.313	0.324	0.523	0.744	0.557	**0.938**
F$_c$	0.45	0.453	0.544	0.748	0.653	**0.938**
RE$_N$	**0.694**	0.558	0.355	0.604	0.556	0.541
PR$_N$	0.42	0.397	0.485	0.656	0.618	**0.967**
F$_N$	0.523	0.464	0.41	0.629	0.585	**0.694**
MMR	**0.613**	0.549	0.404	0.54	0.571	0.587
Detected Cmplx	243	213	88	652	115	NA

As shown in Fig. 3, SVM-Net is able to detect more matched complexes (76 matching complexes, α=30) than other state-of-the-art methods with higher recall and precision.

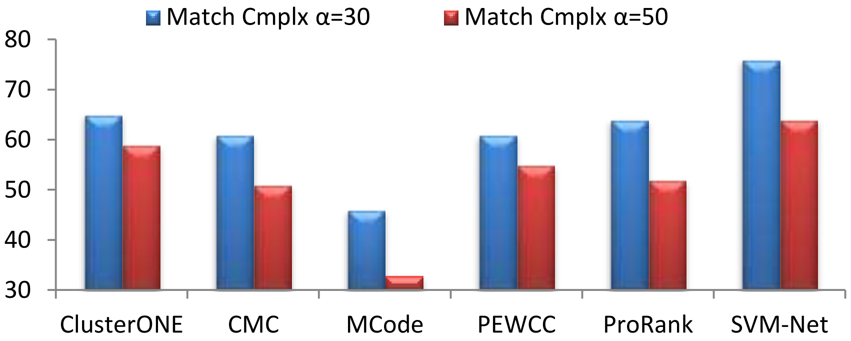

Fig. 3. The number of matched complexes detected by SVM-Net in comparison to ClusterONE, CMC, MCode, PEWCC and ProRank.

4. Conclusion

Most of currently available methods for detecting multi-protein complexes mainly focus on topological information and fail to consider the information from protein primary structure. Protein sequence information is of considerable importance for protein complex detection. Based on this observation, we propose a method called SVM-Net to discover multi-protein complexes from yeast PPI network. SVM-Net extracts valuable features from the protein primary structure (amino acid background frequency) and the topology of the PPI network which is helpful for the effective detection of the multi-protein complex. The experimental works conducted on a PPI network prepared by Gavin et al. (2006)

and a reference dataset of 81 multi-protein complex showed that SVM-Net outperforms five of the state-of-the-art protein complex detection methods. In the future, more valuable features such as gene ontology or gene expression can be incorporated.

Acknowledgements

The authors would like to acknowledge the assistance provided by the College of Information Technology and the National Research Foundation (Grants Ref. No. 21 T021 and 31T052-UPAR) and the Research Support and Sponsored Projects Office at the United Arab Emirates University (UAEU).

References

Acencio M.L., Lemke N. (2009). Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. BMC Bioinformatics, 10, 290.

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

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

Bader G.D., Christopher W.H. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 4, 2.

Chin CH. (2010). “Prediction of Essential Proteins and Functional Modules From Protein-Protein Interaction Networks”. National Central University.

Chiou-Yi H., Chang-Biau Y., Zih-Jie Y., Chiou-Ting T. (2013). Prediction of Protein Essentiality by the Support Vector Machine with Statistical Tests. Evolutionary Bioinformatics, 9, 387–416.

Cristianini N., Shawe-Taylor J. (2000). “An introduction to Support Vector Machines”, Cambridge, UK: Cambridge University Press.

Csárdi G., Nepusz T. (2006). The igraph software package for complex network research. InterJournal.

Dongen S. (2000). “Graph Clustering by Flow Simulation”, University of Utrecht, Netherlands.

Gavin A.C., Aloy P., Grandi P., Krause R., Boesche M., Marzioch M., Rau C., Jensen L.J., Bastuck S., Dümpelfeld B., Edelmann A., Heurtier M.A., Hoffman V., Hoefert C., Klein K., Hudak M., Michon A.M., Schelder M., Schirle M., Remor M., Rudi T., Hooper S., Bauer A., Bouwmeester T., Casari G., Drewes G., Neubauer G., Rick J.M., Kuster B., Bork P., Russell R.B., Superti-Furga G. (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature, 440(7084), 631–636.

Guimei L., Wong L., Chua H. N. (2009) Complex discovery from weighted PPI networks. Bioinformatics, 25(15), 1891–1897.

Hu P., Janga S.C., Babu M., Diaz-Mejia J.J., Butland G., Yang W., Pogoutse O., Guo X., Phanse S., Wong P., Chandran S., Christopoulos C., Nazarians-Armavil A., Nasseri N.K., Musso G., Ali M., Nazemof N., Eroukova V., Golshani A., Paccanaro A., Greenblatt J.F., Moreno-Hagelsieb G., Emili A. (2009). Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS Biol., 7(4), e96.

Huynen M., Snel B., Lathe W., Bork P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res., 10, 1204-1210.

Hwang Y.C., Lin C.C., Chang J.Y., Mori H., Juan H.F., Huang H.C. (2009). Predicting essential genes based on network and sequence analysis. Mol Biosyst., 5(12), 1672–1678.

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

Lin C.Y., Chin C.H., Wu H.H., Chen S.H., Ho C.W., Ko M.T. (2008). Hubba: hub objects analyzer—a framework of interactome hubs identification for network biology. Nucleic Acids Res., 36, 438–43.

Lin C.Y., Yang C.B., Hor C.Y., Huang K.S. (2010). Disulfide bonding state prediction with SVM based on protein types. “Proc IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Appl.”, Changsha, China, Sep 23-26, pp. 1436-1442.

Mark H., Eibe F., Geoffrey H., Pfahringer B., Reutemann P., Witten I.H. (2009). The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter, 11(1), 10-18.

Mewes H.W., Frishman D., Güldener U., Mannhaupt G., Mayer K., Mokrejs M., Morgenstern B., Münsterkötter M., Rudd S., Weil B. (2002). Mips: a database for genomes and protein sequences. Nuc. Acids Res., 30(1), 31-34.

Nepusz T., Yu H., Paccanaro A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods, 9, 471–472.

Rong-en F., Pai-hsuen C., Chih-jen L. (2005). Working set selection using second order information for training SVM. Journal of Machine Learning Research, 6, 1889-1918.

Simonis N., Helden J., Cohen G.N., Wodak S.J. (2004). Transcriptional regulation of protein complexes in yeast. Genome Biology, 5, R33.

Vanunu O., Magger O., Ruppin E., Shlomi T. Sharan R. (2010). Associating Genes and Protein Complexes with Disease via Network Propagation. PLoS Comput. Biol., 6, 1.

Vapnik V.N. (1998). “Statistical Learning Theory”, Wiley.

Zaki N.M., Berengueres J., Efimov D. (2012). Detection of protein complexes using a protein ranking algorithm. Proteins: Structure, Function, and Bioinformatics, 80(10), 2459-2468.

Zaki N.M., Bouktif S., Lazarova-Molnar S. (2011). A Combination of Compositional Index and Genetic Algorithm for Predicting Transmembrane Helical Segments. PLoS ONE 6(7), e21821.Brohee S., van Helden J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics, 7, 488.

Zaki N.M., Dmitry D., Berengueres J. (2013) Protein Complex Detection using Interaction Reliability Assessment and Weighted Clustering Coefficient. BMC Bioinformatics, 14, 163.