# Classification of Yeast Protein Complexes Using Topological and Physicochemcial Features

Szu-Yu Chou[1], Chien-Hung Huang[2]*
Department of Computer Science and Information Engineering, National Formosa University
Hu-wei, Yun-Lin, Taiwan 632.
* Corresponding author
[1] darkkiss9527@hotmail.com , [2] *chhuang@nfu.edu.tw

Ka-Lok Ng*
Department of Biomedical Informatics, Asia University,
Wufeng Shiang, Taichung, Taiwan 41354.
* Corresponding author
ppiddi@gmail.com

***Abstract*—It is known that protein complexes play important roles in many cellular processes. Many protein complex prediction methods are based on the identification of pseudo-cliques, and dense protein-protein interaction (PPI) regions. Amino acids' physiochemical properties are not generally used in the feature representation of protein complexes. The results of the physiochemical properties study for yeast protein complexes are reported. A 10-fold cross-validation test is performed based on the features to test the classification accuracy of the Support Vector Machine (SVM). It is found that the physiochemical properties serve as additional important features besides PPI and Gene Ontology.**

***Index Terms- protein complexes, physiochemical properties, principle component analysis, logistic regression, support vector machine***

## INTRODUCTION

Proteins perform distinct and well-defined functions. In the last ten years or so, we have seen many high-throughput experimental techniques attempt to crystallize proteins [1], determine peptide sequences using mass spectrometry [2-4] , three-dimensional structures, and protein-protein interaction networks (PPIN) through, such as, NMR spectroscopy [5-6] and the genome-wide yeast two hybrid [7]. A lot is known about how proteins interact at the cellular level. There are a few major publicly available databases on PPI, which include DIP [8-9] (Xenarios et al. 2001, Salwinski et al. 2004), BOND [10] (Alfarano et al. 2006), HPRD [11] (Mishra et al. 2006), IntAct [12] (Kerrien et al. 2007), MINT [13] (Chatraryamontri et al. 2007), MIPS [14] (Mewes et al. 2008) and BioGRID [15] (Breitkreutz et al. 2008).

It was reported that [16] in the yeast organism, the protein-protein interactions (PPI) are not random but well organized. It was also found that most of the neighbors of highly connected proteins have few neighbors, that is highly connected proteins are unlikely to interact with each other.

It is known that proteins do not work alone; rather they act in a cooperative manner through the formation of protein complex in many biological processes [17]. Yeast is chosen as the model system for the present study for three reasons; (i) the complete genome sequence has been completed since year 1996 [18], (ii) the PPI are well studied [19], and (iii) there are many molecular tools, such as cDNA, genomic libraries, bacterial artificial chromosomes, microarray[20], and ESTs are available for the biological functions study. Furthermore, various approaches, such as genomic datasets [21] or the mass spectrometry scores and PPI data [22], are used to make yeast protein complexes prediction.

Recent experimental studies indicate that the protein complex can be visualized as a unit composed of the cores, modules and attachments [23-24]. Core proteins are proteins that have relatively more interactions among themselves and belong to a unique protein complex [25-26]. Attachment proteins bind to the core proteins with relative fewer interactions among them. Module proteins are a subset of the attachment, which are always present together, and module proteins can be present in more than one complex. A recent study indicated that the prediction of protein complexes based on the core-attachment model can achieve better performance than graphical approaches [27].

It is known that many protein complex prediction calculations are based on the identification of pseudo-cliques [28-31]. Our previous study indicated that interaction dense regions represent yeast protein complexes in up to 20% of cases [32]. A rather large proportion of protein complexes have a lower density of PPI. It is conjectured that prediction approaches based on the assumption that complexes are composed of highly PPI dense regions can predict a rather limited number of complexes.

In this study we propose characterizing the protein complexes by considering physiochemical properties. Amino

acids' physiochemical properties are not generally used in the feature representation of protein complexes. There is an attempt in using physiochemical properties in detecting remote protein homology [33] with rather successive performance. Here we propose to consider the following physiochemical properties of a complex; i.e. hydrophobic, hydrophilic, pI value (iso-electric point), half-life, length distribution and amino acid compositions. Instead of trying to classify complexes from PPI data only, one of the major objectives of the present study is to classify complexes based on physiochemical properties as well as PPI information.

It is proposed that the results of this work are helpful in improving the prediction accuracy whether a complex predicted by PPI is a real complex or not; assuming that physiochemical parameters are supplied. In a previous study [34] it is suggested that physiochemical properties may serve as additional important features in classifying yeast protein complexes.

Physiochemical parameters are used to construct the feature vectors and trained by support vector machine (SVM) [35] for protein complexes classification purpose. A 6-fold cross-validation test is performed to validate the classification accuracy of SVM based on the major features.

## METHODS

A total of 1643 and 491 yeast protein complexes data were retrieved from BOND [36] and Yeast [37] respectively. All the subunits' accession numbers are labeled according to gene index for protein.

### *A. Protien-protein interaction density*

The first parameter is called the density of interaction, which describes the experimentally recorded PPI among the subunits of a protein complex relative to the maximum possible PPI (i.e. clique). Given that a protein complex has $N$ subunits, it can have $N*(N+1)/2$ possible PPIs, including self-interaction. Then the density of PPI, $\rho$, among the subunits of a protein complex, is given by

$$\rho = \frac{2s}{N*(N+1)} * 100\% \cdot \quad (1)$$

where $s$ is the observed number of PPIs among the subunits, and the fraction of nodes in the largest connected cluster, $\lambda$, so-called connectedness, of a protein complex is defined by,

$$\lambda = \frac{M+1}{N} \quad (2)$$

where $M$ is the largest distance of the largest connected cluster obtained by the Floyd-Warshall shortest path algorithm. For $M$ equals to zero, $\lambda$ is set to zero, this simply means no subunit is connected. PPI data are obtained from the BioGrid database.

### *B. Sequence similarity*

An all-against-all pairwise sequence alignment is performed by using the BLAST program. Output files reported by the BLAST program for the whole yeast complexes are parsed, and the average of the percentage of similarity value for each complex is computed. The average of the sequence similarity of a complex $C$, $I_C$, is defined by,

$$I_C = \frac{2}{N(N-1)} \sum_{i<j \in C} I_{ij} \quad (3)$$

where $I_{ij}$ denotes the sequence similarity percentage reported by BLAST, $i$ and $j$ are labels ($i = 1, \ldots N-1$) which denote the complex subunits, and $C$ denotes one of the yeast protein complexes.

It is also noted that special care is required for complexes consisting of repeated subunits. Let $D$ denote a complex composed of $n$ and $k$ distinct and repeated subunits respectively, i.e. $D = \{A, \ldots..N, \alpha_1, \ldots. \alpha_k\}$, where $\alpha_1$, …. and $\alpha_k$ are repeated subunits. For the complex $D$, the sequence similarity percentage, $I$, received contributions from $I^{cross}$ and $I^{repeated}$, which are defined by

$$I^{cross} = \sum_{\substack{i=1,\ldots.n \\ \alpha=1,\ldots.k}} I_{i\alpha} \quad (4)$$

and $I^{repeated}$ equals to $k(k-1)/2$. The average of the sequence similarity of $D$ is given by

$$I_D = \frac{1}{C_2^{N+k}} [\sum_{i<j=2,..N} I_{ij} + I^{cross} + I^{repeated}] \quad (5)$$

The same argument can be generalized for more than one type of repeated subunits. In order to demonstrate the statistical significance of the BLAST results, a randomize version of the protein complexes is performed, in which subunits are randomly assigned to each complex, while keeping the total number of protein complexes, subunits, and the number of subunits within each complex the same as the original. The randomized results are compared with the raw results.

### *C. Jaccard index of Gene Ontology*

Given the Gene Onotology (GO) annotation of protein complexes $A$ and $B$, the Jaccard index (JI) of GO is defined as

$$JI = \frac{|A \cap B|}{|A \cup B|} \quad (6)$$

where $|A \cap B|$ and $|A \cup B|$ denote the cardinality of $A \cap B$ and $A \cup B$ respectively.

### *D. Physiochemical properties*

Values of the physiochemical properties for the subunits can be computed from several bioinformatics tools; see Table 1.

The hydrophobic and hydrophilic values for the twenty types of amino acids are provided by the Kyte and Doolittle (KD) scale [38]. ExPASy provides the tool, ProtParam [39], to compute the physicochemical parameters of a protein sequence, such as pI, half-life, length distribution of subunits within a complex. ProtParam computes various physicochemical properties that can be deduced from the protein sequence.

TABLE I. TOOLS USED FOR COMPUTING THE PHYSIOCHEMICAL PROPERTY OF A PROTEIN SUBUNIT

| Physiochemical property | Tool |
|---|---|
| hydrophobic, hydrophilic | KD scale [38] |
| pI value | ProtParam [39] |
| half-life | ProtParam |

The half-life is a prediction of the time it takes for half of the amount of protein in a cell to disappear after its synthesis in the cell. ProtParam relies on the "N-end rule", which relates the half-life of a protein to the identity of its N-terminal residue [40]. The length and amino acid compositions are obtained from the subunit's sequence information.

Since the physiochemical value varies from subunit to subunit, the coefficient of variation (*CV*) is introduced to represent the whole complex. Assuming a protein complex *D* contains *n* subunits, given that subunit *i* has a physiochemical value $x_i$, let *CV(D)* represents the coefficient of variation for physiochemical value *x* for the whole complex, and it is defined by

$$CV(D) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \Big/ \sum_{i=1}^{n} x_i \qquad (7)$$

where $\bar{x}$ denotes the average of the physiochemical value for the whole complex.

Machine learning method SVM is used to train the input feature vectors. In particular, the LIBSVM [41] is used for complex classification in our study. LIBSVM provide several kernel function for classification, i.e. linear, polynomial, sigmoid and radial basis function (RBF).

Randomized samples are generated in order to train protein complex classification using SVM. Tests are performed in which assignment of protein subunits is randomized while keeping seven subunits for each complex in forming random complexes.

Since the random set will resulted in trivial classification, therefore, we made use of the BioGrid data, input that in COACH [31] to generated a set of 542 pseudo-complexes. COACH is a protein complex prediction method which infers complexes using graph clustering techniques. This set of pseudo-complexes is put together with the random set to form the complete randomized set. Then, the physiochemical values for each of the complex in the randomized set are computed. The results of the randomized set are taken together with the original complexes' values to form the training set, and input into the SVM for training.

## RESULTS

### *E. Protien-protein interaction density*

Figure 1 showed that around 45% of the protein complexes obtained from BOND has a density of interaction of over 90%, and the other 55% of complexes account for other $\rho$ values. It is also a surprise that more than 90% (yeast data) of protein complexes have $\rho$ values less than 30%.

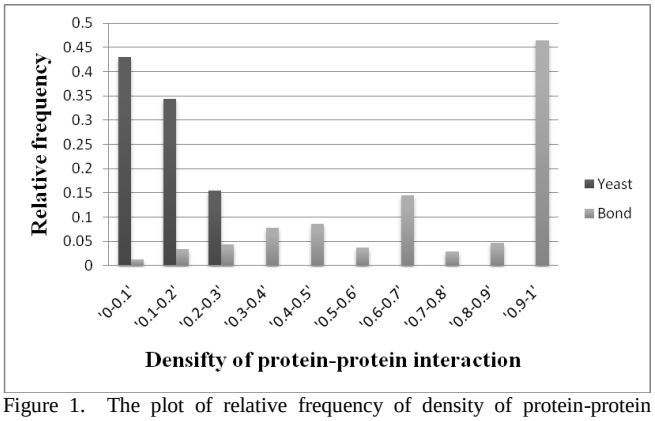


Figure 1. The plot of relative frequency of density of protein-protein interaction of yeast protein conplexes obtained from the Yeast and BOND databases.

In other words, quite a significant number of complexes do not have PPI among their subunits. These results suggested that algorithms based on the assumption that complexes are composed of highly PPI dense regions can only predict a limited numbers of complexes.

### *F. Seqeunce similarity*

For the protein complex subunits sequence similarity study, the relative frequency of protein complexes versus the sequence identity and similarity percentage intervals is depicted in Figure 2 and 3 respectively. It is noted that the distributions of the relative frequency of protein complexes against the sequence identity and similarity interval are not uniform. Most of the protein complexes have an average sequence identity and similarity in the 30% -40% and 50%-60% intervals respectively.

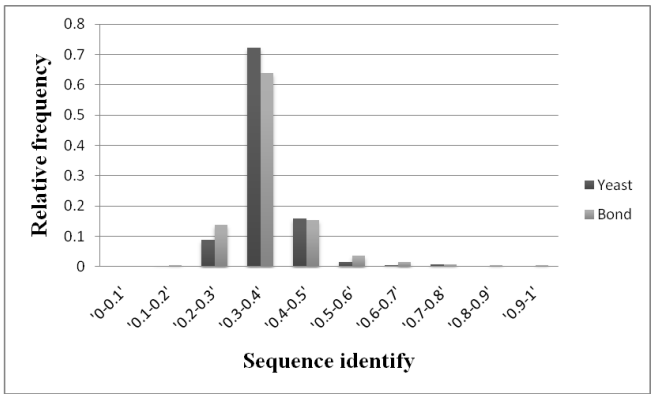


Figure 2. The plot of relative frequency of sequence identity of yeast protein conplexes obtained from the Yeast and BOND databases.

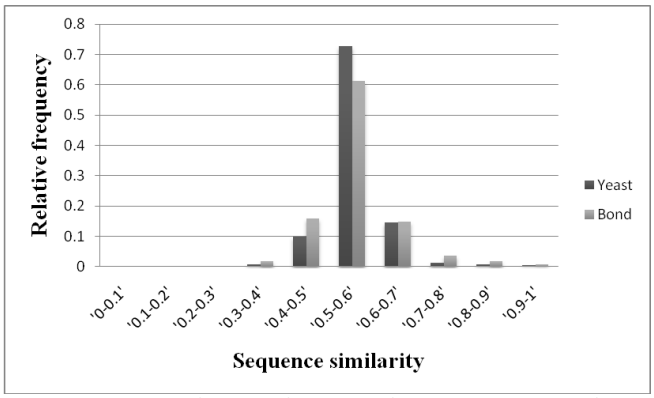


Figure 3. The plot of relative frequency of sequence similarity of yeast protein conplexes obtained from the Yeast and BOND databases.

If it is assumed that the "twilight zone" of the sequence similarity for two protein sequences is taken to be 30%, then all of the complexes have an overall average sequence similarity over the 30% threshold. In other words, protein complexes are composed of subunits with similar protein sequences. Furthermore, both of the yeast and BOND data show similar distributions of relative frequency.

## *G. Gene ontology*

In Figure 4, we plot the relative frequency of Jaccard index of GO for all the yeast protein complexes. It is found that about 28.9% and 17.5% of the complexes have Jaccard index over 0.3 for the BOND and yeast database respectively. This suggests that protein complexes are composed of subunits with a fair amount of similar molecular functions. It is noted that both of the yeast and BOND data show similar distributions of relative frequency.

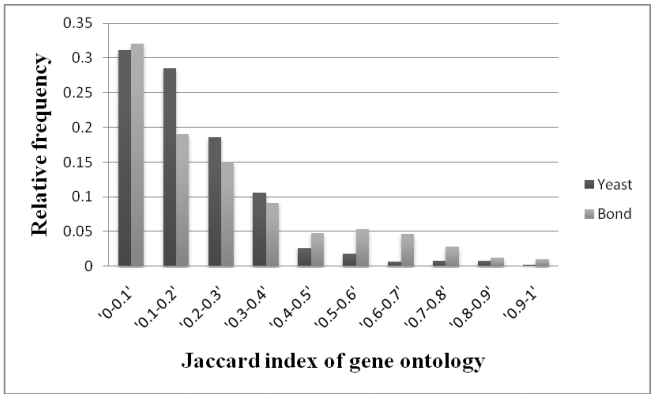


Figure 4. The plot of relative frequency of Jaccard index of gene ontology of yeast protein conplexes obtained from the Yeast and BOND databases.

## *H. Hydrophobic and hydrophilic properties*

The hydrophobic ($H_b$) and hydrophilic ($H_p$) values for the amino acids in each subunit are added together. Since both the $H_b$ and $H_p$ values vary from subunit to subunit within a complex, therefore, their averages are computed, and these two averages are used to represent the whole complex.

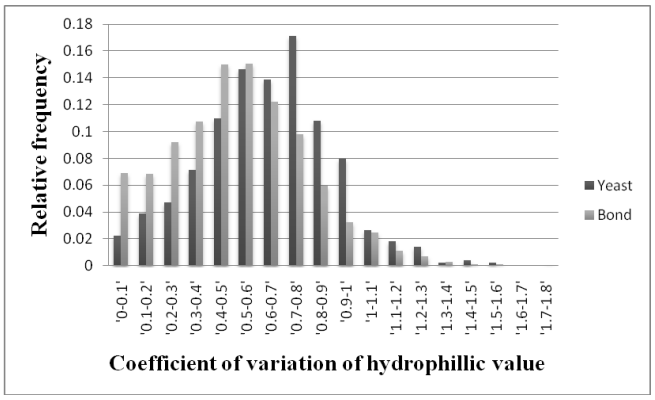


Figure 5. The plot of relative frequecny of CV of hydrophilic for yeast protein conplexes obtained from the Yeast and BOND databases.

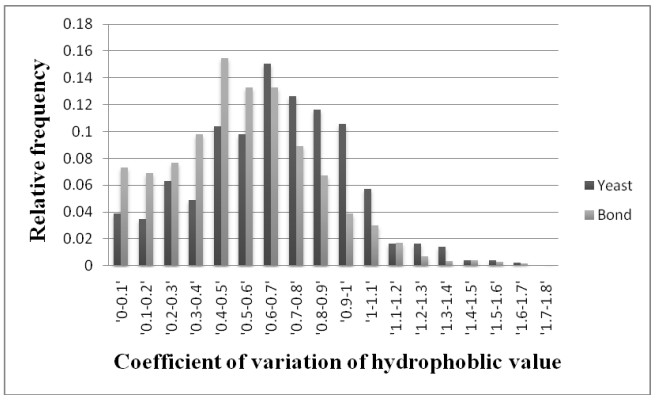


Figure 6. The plot of relative frequecny of CV of hydrophobic for yeast protein conplexes obtained from the Yeast and BOND databases.

It is noted that the BOND data shows a smaller CV hydrophilic and hydrophobic value relative to the yeast database.

## *I. pI value*

The pI value for every protein complex's subunit is obtained from Protoparm. For each complex, the *CV* for pI is computed, after that we grouped the values *CV* into a 10% interval. And the plot of the relative frequency versus the 10% interval is shown in Figure 7. It is noted that the BOND data shows a smaller CV pI value relative to the yeast database.

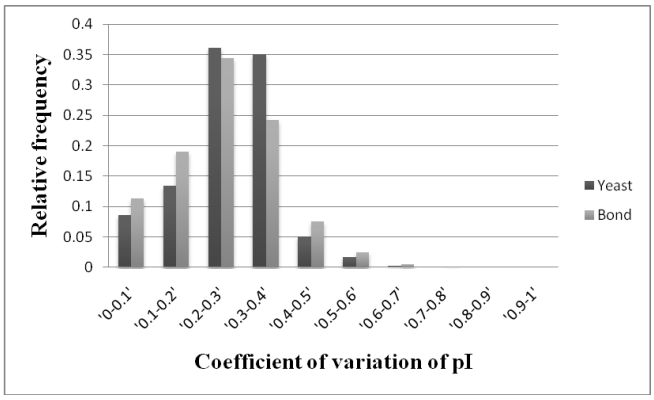


Figure 7. The plot of relative frequecny of CV of pI for yeast protein conplexes obtained from the Yeast and BOND databases.

## J. *Length distribution of subunits within a complex*

Length of every protein subunits for the complexes is computed. For each complex, the *CV* is computed, then we group the *CV* into a 10% interval, the plot of the relative frequency versus the 10% interval is shown in Figure 8. Again, it is noted that the BOND data shows a smaller CV value relative to the yeast database.

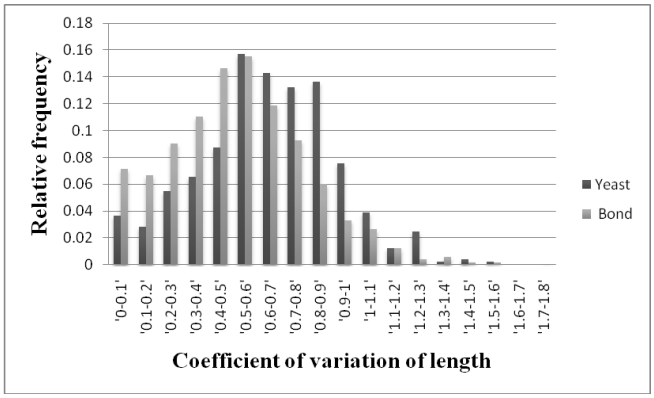


Figure 8. The plot of relative frequecny of CV of length of yeast protein conplexes obtained from the Yeast and BOND databases.

## K. *Support vector machine classification results*

Using the RBF kernel, the classification results predicted by the SVM for several combinations of the major features are given in Table I.

TABLE I. CLASSFICATION RESULTS OF PROTEIN CO MPLEX FOR DIFFERENT FEATURES AND PHYSIOCHEMICAL PROPERTY COMBINATIONS

| Classification features | Accuracy(%) | |
|---|---|---|
| | Yeast | BOND |
| GO | 84.4 | 79.5 |
| bit score | 65.9 | 64.1 |
| Density of PPI | 84.5 | 96.1 |
| *CV* of pI | 64.0 | 58.1 |
| hydrophilic | 52.5 | 58.0 |
| *CV* of length | 53.4 | 57.8 |
| GO & Density of PPI | 89.6 | 96.1 |
| GO & **bit score** | 86.4 | 79.5 |
| Density of PPI & **bit score** | 84.9 | 95.7 |
| GO & *CV* of pI | 84.5 | 79.6 |
| GO & hydrophilic | 84.4 | 79.5 |
| GO & *CV* of length | 84.9 | 79.5 |
| Density of PPI & *CV* of pI | 83.7 | 96.1 |
| Density of PPI & hydrophilic | 84.4 | 96.1 |
| Density of PPI & *CV* of length | 84.6 | 96.1 |
| GO, Density of PPI & **bit score** | 88.8 | 95.8 |
| GO, Density of PPI & pI | 89.0 | 96.2 |
| GO, Density of PPI & hydrophilic | 89.4 | 96.2 |
| GO, Density of PPI, **bit score** & pI | 89.0 | 95.9 |
| GO, Density of PPI, **bit score** & hydrophilic | 88.6 | 95.8 |
| GO, Density of PPI, **bit score**, pI & length | 88.5 | 95.8 |

The results indicate that classification based only on single feature can achieve a 63.9-80.7% accuracy, with GO and pI obtain the highest accuracy for topological parameter and physiochemical feature respectively. With the addition of a second feature, the classification accuracy raises to 69.8-82.6%. Classification accuracy remains more or less over 83% on adding a third feature, where the three features; (i) GO, density of PPI & sequence similarity and (ii) GO, density of PPI & CV of pI achieves the highest (84.8%) and the second highest (84.2%) accuracy respectively. It is found that the highest accuracy (85.1%) is achieved for the combination of five features, i.e. GO, density of PPI, sequence similarity, pI and length.

## SUMMARY

Several topological features and physiochemcial parameters are combined to describe protein complexes and gain better insights for our understanding of protein complex architecture.

Given the physiochemical features for a predicted complex, a possible application of the present study is on improving the accuracy of determining whether the complex inferred by PPI is a real complex or not. The present work can be an interesting discovery that strongly suggests integrating physiochemical data to improve protein complexes prediction.

## ACKNOWLEDGMENT

The work of Chien-Hung Huang is supported by the National Science Council (NSC) of Taiwan, under the grant of NSC 100-2221-E-150-069. The work of Ka-Lok Ng is supported by the NSC of Taiwan, under the grants of NSC 100-2221-E-468 -013 and NSC 99-2221-E-468-016-MY2.

## REFERENCES

[1] C Stevens Raymond, "High-throughput protein crystallization," Current Opinion in Structural Biology, Vol. 10(5), pp.558-563, 2000.

[2] K Gevaert and J.Vandekerckhove, "Protein identification methods in proteomicsElectrophoresis," Vol. 21(6), pp.1145-54, 2000.

[3] Dayin Lin, David L. Tabb and John R. Yates, "III Large-scale protein identification using mass spectrometry, " Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics, Vol. 1646(1-2), 21, pp.1-10,2003.

[4] Ruedi Aebersold and Matthias Mann. "Mass spectrometry-based proteomics," Nature Vol. 422, pp.198-207,2003.

[5] Liu et al., "NMR data collection and analysis protocol for high-throughput protein structure determination," Proceedings of the National Academy of Sciences, Vol. 102(30), pp.10487-10492 ,2005.

[6] Udo Heinemann, "Gerd Illing and Hartmut Oschkinat High-throughput three-dimensional protein structure determination," Current Opinion in Biotechnology, Vol. 12(4), pp.348-354,2001.

[7] Charlotte M. Deane, Łukasz Salwin´ ski, Ioannis Xenarios, and David Eisenberg, "Protein Interactions: TWO METHODS FOR ASSESSMENT OF THE RELIABILITY OF HIGH THROUGHPUT OBSERVATIONS," Molecular & Cellular Proteomics Vol. 1, pp.349-356,2002.

[8] I.Xenarios, L.Salwinski, X.J.Duan, P.Higney, S.M.Kim and D Eisenberg. "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Res*., Vol. 30, pp.303-305,2002.

[9] L. Salwinski, C.S.Miller, A.J.Smith, F.K.,Pettit J.U Bowie., and D. Eisenberg. "The Database of Interacting Proteins," *Nucl. Acids Res*. Vol. 32, pp.D449-D451,2004.

[10] Alfarano et al. "The Biomolecular Interaction Network Database and related tools 2005 update," Nucleic Acids Res. Vol.33 (Database issue):pp.D418-D424,2006.

[11] Mishra et al. "Human protein reference database - 2006 update," *Nucleic Acids Res*., Vol. 34: pp.D411-D414,2006.

[12] Kerrien S. et al. "IntAct - open source resource for molecular interaction database," Nucleic Acids Res., Vol. 35: pp.D561-D565,2007.

[13] Chatraryamontri Andrew, Ceol Arnaud, Palazzi Luisa Montecchi, Nardelli Giuliano, Schneider Maria Victoria, Castagnoli Luisa, and Cesareni Gianni. "MINT: the Molecular INTeraction database," *Nucleic Acids Res.*, Vol. 35: pp.D572-D574,2007.

[14] Mewes et al. "MIPS: analysis and annotation of genome information in 2007," *Nucleic Acids Res*., Vol. 36, pp.D196-D201,2008.

[15] Breitkreutz et al. "The BioGRID Interaction Database: 2008 update," *Nucleic Acids Res*., Vol. 36, pp.D637-D640,2008.

[16] S Maslov, and K.Sneppen. "Specificity and stability in topology of protein networks," *Science*, Vol. 296, pp.910-913,2002.

[17] C. Kleanthous. "Protein-protein Recognition," Oxford University Press,2000.

[18] Goffeau et al., Life with 6000 genes. Science, Vol. 274(5287):pp.546, 563-7,1996.

[19] Ito et al., "Toward a protein–protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins," Proceedings of the National Academy of Sciences, Vol. 97(3), pp.1143-1147,2000.

[20] EA Winzeler, B Lee, JH McCusker, RW Davis. "Whole genome genetic-typing in yeast using high-density oligonucleotide arrays," Parasitology.; Vol. 118 Suppl:pp.S73-80,1999.

[21] R Jansen, N Lan, J Qian, M Gerstein. "Integration of genomic datasets to predict protein complexes in yeast." *J Struct Funct Genomics* Vol. 2: pp.71-81,2002.

[22] Krogan et al., "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," Nature Vol. 440, pp.637-643,2006.

[23] Z. Lubovac, J. Gamalielsson and B. Olsson, "Combining Functional and Topological Properties to Identify Core modules in Protein Interaction Networks,". *Proteins: Structure, Function, and Bioinformatics*, Vol. 64(4), pp. 948-959,2006.

[24] Pang et al., "Are Protein Complexes Made of Cores, Modules and Attachments?" *Proteomics*, Vol. 8(3), pp. 425-434,2008.

[25] Gavin et al., "Proteome Survey Reveals Modularity of the Yeast Cell Machinery," *Nature*, Vol. 440 , pp. 631–636, 2006.

[26] Zhang et al., "From Pull-down Data to Protein Interaction Networks and Complexes with Biological Relevance," *Bioinformatics*, Vol. 24(7), pp. 979-986, 2008.

[27] Leung et al., "Predicting Protein Complexes from PPI data: a Core-attachment Approach," *Journal of Computiponal Biology*, Vol. 16(2), pp. 133-144, 2009.

[28] V.Spirin and L.A Mirny, "Protein Complexes and Functional Modules in Molecular Networks," *Proc. N.atl. Acad. Sci.*, Vol. 100(21):pp.12123-12128, 2003.

[29] Altaf-Ul-Amin et al.:"Development and Implementation of an Algorithm for Detection of Protein Complexes in Large Interaction Networks,"*BMC Bioinformatics*,7, 207:pp.1-13,2006.

[30] Yu et al.: "Predicting Interactions in Protein Networks by Completing Defective Cliques," *Bioinformatics*, Vol. 22(7): pp.823-829, 2006.

[31] Xiaoli Li, Min Wu, Chee-Keong Kwoh and See-Kiong Ng, "Computational approaches for detecting protein complexes from protein interaction networks: a survey," *BMC Genomics*, Vol. 11(suppl 1), S3,2010.

[32] Chien-Hung Huang, K-.T Chao and Ka-Lok Ng, "Protein Complexes Subunits Interaction Topology and Sequence Identity," *The 3rd International Conference on Computer Research and Development (ICCRD 2011)*, Shanghai, China, Mar. pp.11-13, 2011.

[33] Yuchen Yang, Trantoso Erwin and Li Kuo-Bin. "Remote protein homology detection using recurrence quantification analysis and amino acid physiochemical properties," *Journal Theoretical Biology*, Vol. 252, pp. 145-154.Ingo Steinwart and Andreas Christmann, Support Vector Machines, 1st edition, Springer, 2008

[34] Chien-Hung Huang, K-.T Chao and Ka-Lok Ng. "Physiochemcial Features Study of Protein Complexes," *IEEE ICCSIT 2011*, Chengdu, China, Jun. pp.10-12, 2011

[35] Ingo Steinwart and Andreas Christmann, Support Vector Machines, 1st edition, Springer, 2008.

[36] Alfarano et al., "The Biomolecular Interaction Network Database and related tools 2005 update," Nucleic Acids Res., Vol. 33, pp.D418-24,2005.

[37] http://yeast-complexes.russelllab.org/

[38] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal Molecular Biology*, vol. 157, 1982, pp. 105-132,1982.

[39] Gasteiger et al., "Protein Identification and Analysis Tools on the ExPASy Server," in John M. Walker (ed): *The Proteomics Protocols Handbook*, Humana Press, pp. 571-607,2005.

[40] http://tw.expasy.org/tools/protparam-doc.html

[41] http://www.csie.ntu.edu.tw/~cjlin/libsvm/