Prediction of Protein-Protein Interactions Using Bayesian Networks

Thanh Phuong Nguyen† **Tu Bao Ho†** **Ngoc Binh Nguyen‡**

†School of Knowledge Science, Japan Advanced Institute of Science and Technology, Japan

‡College of Technology, Vietnam National University - Hanoi, Vietnam

{phuong,bao}@jaist.ac.jp, nnbinh@vnu.edu.vn

Abstract

Interactions among proteins are essential to almost all cellular functions. Several experimental and computational methods have been proposed to discover protein-protein interactions (PPI). However the reliability of these interactions is still challenging. Our motivation is to more reliably predict protein-protein interactions by mining and combining multiple biologically significant data. By integrating function catalog and biological process data, especially domain fusion data in Bayesian networks, a considerable number of putative PPI was predicted. In addition to the large number of predicted PPI, the main contribution of our work is predicting more reliable PPI by combining both protein information and domain information. Experimental results demonstrated the reliability of predicted PPI in three main points: (i) the large number of overlapped interactions between predicted PPI set and two well-known PPI databases, (ii) the large number of domain-domain interactions covered by predicted PPI, and (iii) the physical properties of predicted PPI.

Keywords: protein-protein interactions, Bayesian networks, biological processes, functional catalogs, domain fusion.

1 Introduction

Proteins are macro molecules made of twenty amino acids arranged in a linear chain, which participate in every process within cells. Protein domains (domains in short) are the key elements in proteins. Different domains in different proteins sometimes fused in other proteins to perform specific biological functions in particular environment conditions. This phenomenon is called domain fusion.

In genome era, more and more genomes and their encoded proteins are successfully studied. However, simply knowing the list of genes and proteins is not sufficient to tell us about the complex biological functions in cell, and most of proteins in cell are considered not to be independent individuals. They could interact permanently or transiently with the others to function in cell. Protein-protein interactions are at the heart of biological activities

Protein-protein interaction prediction is one of the most important problems in biology. With much effort of biological scientists, protein-protein interactions were early discovered from many experimental methods such as yeast two-hybrid, phage display, affinity purification and mass spectrometry, and protein micro-arrays [17], [20]. Until recently, the results from these expensive experimental methods are little overlapped.

Since the constantly increasing numbers of published biological databases, there are a multitude of computational methods to detect protein-protein interactions. Some work tried to propose the new computational techniques to effectively infer PPI [1], [4], [6], [12], [15]. From a different point of view, other work have recently been tried to look for biologically significant data related to PPI, and then combined them in computational frameworks [3], [8], [10], [13]. Owing to the combination of multiple databases, PPI can be better predicted in a comprehensive way. Though domains are basic parts in proteins [11], most of previous works did not use this appealing information in their integrative frameworks.

In this paper, our purpose is to more reliably predict protein-protein interaction by integrating domain information, in particular domain fusion, using Bayesian network framework. Bayesian networks have been shown to be powerful [8]. We took an effort to improve this framework by integrating domain fusion information in addition to function catalog and biological process. As a

result, our contribution is proposing of a method which uses not only protein information but also domain information. Studying on the deeper mechanism of protein-protein interaction, our method is promising to more reliably predict PPI.

In experiments, among 7 millions of yeast protein pairs (matched from the three genomic data sources), we predicted a considerable number of protein-protein interactions. We evaluated the reliability of predicted interactions in three ways. First, comparing with two well-known protein interaction databases, Comprehensive Yeast Genome Database [24] and Database of Interacting Proteins [25], there are much overlapped between the predicted interaction set and these databases. Second, we extracted domain-domain interactions from InterDom database [11], and checked whether protein interactions had domain-domain interactions. Because domain-domain interactions (DDI) are biological mechanisms underlying PPI, the large number of covered DDI persuaded the reliability of predicted PPI. Third, the reliability of putative PPI was reconfirmed by the physical properties.

For further study on protein complexes and cellular pathways, we discovered a large number of triplets (two interacting couples sharing one protein). These strong triplets may be building blocks to form various biological pathways and functional complexes.

In Section 2, we present our proposed methods and materials. The experimental results and evaluation will be shown Section 3. Finally, we make some conclusions and discuss some future work in Section 4.

2 Materials and Methods

In this part, we present our method integrating three genomic data sources using Bayesian networks to predict protein-protein interactions. We first extracted complexes catalog from MIPS database and the yeast localization data [22] as positive and negative training data sets, respectively. Then MIPS functional catalogs [24], GO biological processes [21], and domain fusion [23] are the three genomic features. These features are combined to train the Bayesian networks. Figure 1 presents the proposed framework.

2.1 Generating Positive and Negative Training Datasets

Two proteins are in the same complex are known to be likely interact to each other [8]. The positive examples were chosen from the MIPS complex database. In this database, proteins for the yeast species *Saccharomyces cerevisiae*, are categorized into various hierarchy classes, such as *intracellular transport complexes* having two sub-catalogs: *clathrin and clathrin-associated protein (AP) complex*. Similar to [8], only classes containing single complexes are considered to extract positive examples. The total of positives is 8,250 protein pairs.

Until now, there is not any available database for non-interacting protein pairs. There are two popular ways to generate negative examples in studies on protein interactions [2]. The first one is randomly choosing the protein pairs that are not included in the protein interaction set (the positive set). The second one is based on an assumption that if two proteins belong to different compartments in the cell, they are likely to have no chance to interact with each other [2]. We then followed the second approach to generate negative examples. From yeast localization data [24], 2,708,746 protein pairs are considered as negatives in our experiments.

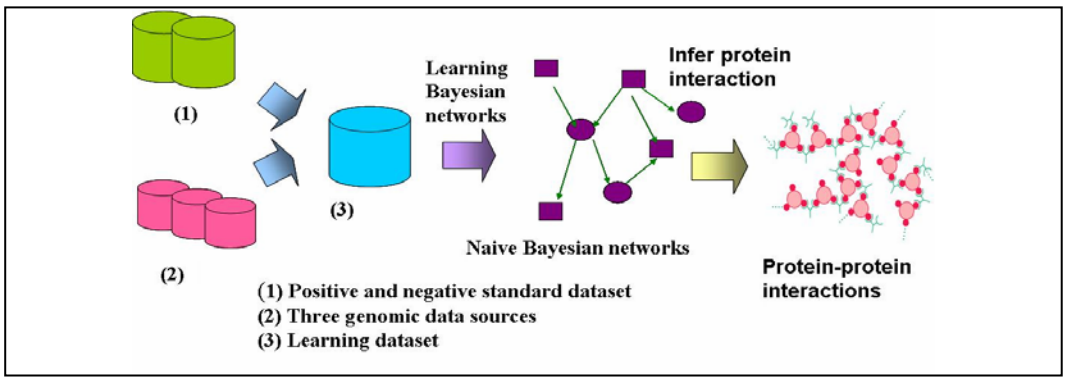


Figure 1. The Bayesian framework for predicting protein-protein interactions.

2.2 Extracting Domain Fusion, Function Catalog and Biological Process Data

Three genomic features, domain fusion, functional catalog and biological process, are combined in our work. Because these features have the biologically close associations with PPI, they are useful features for predicting PPI.

Two proteins functioning in the same biological process are more likely to interact than two proteins involved in different processes. Thus, we can infer PPI from function catalog and biological process data. We collected information from two catalogs of protein functional information: MIPS (Munich Information Center for Protein Sequences) functional catalog [22], and GO (Gene Ontology) biological processes [21].

In addition, domains of interacting proteins can fuse in other proteins to perform some specific functions. It is said that domain fusion and protein interactions have strong relations. If domains of protein A and domains of protein B fuse in protein C, we can predict an interaction between two proteins, A and B [17], [18]. Domain fusion data is referred from Domain Fusion Database [18], [23]. The purpose of the work in [17] is to discover domain fusion using relational algebra and sequence data. Because of the high accuracy when validating with PPI data, we chose domain fusion data from these data sources. The Bayesian network approach can be more effective with predictive features [17], [9]. After investigating various genomic features, three extracted features are considered to be useful for predicting protein-protein interactions.

We used the same procedure proposed in [8] to quantify the functional similarity between two proteins. Protein pairs are binned to five intervals (Inv1, Inv2, Inv3, Inv4, and Inv5) according to their similarity. We have five feature values for MIPS function and GO process features, i.e. Inv1 (1-9), Inv2 (10-99), Inv3 (100-1000), Inv4 (1000-10000), and Inv5 (10000 - infinite). For domain fusion feature, there are two feature values, i.e. "yes" (if having domain fusion), and "no" (if not having domain fusion). All proteins are identified by their ORF (Open Reading Frame).

Table 1. The parameters of the naïve Bayesian network.

Go process							
Feature value	**pos**	**neg**	**Sum(pos)**	**Sum(neg)**	**P(GO process/ pos)**	**P(GO process/ neg)**	***GO_L***
Inv1	98	825	98	825	0.01261099	0.001278856	9.861152
Inv2	779	3336	877	4161	0.100244499	0.005171227	19.38505
Inv3	525	10242	1402	14403	0.067558873	0.015876411	4.255299
Inv4	1005	28251	2407	42654	0.129326985	0.043792667	2.953165
Inv5	5364	602454	7771	645108	0.690258654	0.933880839	0.739129
Sum	***7771***	***645108***					
MIPS function							
Feature value	**pos**	**neg**	**Sum(pos)**	**Sum(neg)**	**P(MIPS function/pos)**	**P(MIPS function / neg)**	***MIPS_L***
Inv1	165	1024	165	1024	0.019666269	0.000779479	25.23002
Inv2	697	4265	862	5289	0.083075089	0.00324656	25.58865
Inv3	741	13119	1603	18408	0.088319428	0.009986313	8.844047
Inv4	6221	47135	7824	65543	0.74147795	0.035879631	20.66571
Inv5	566	1248155	8390	1313698	0.067461263	0.950108016	0.071004
Sum	***8390***	***1313698***					
Domain fusion							
Feature value	**pos**	**neg**	**Sum(pos)**	**Sum(neg)**	**P(domain fusion/pos)**	**P(domain fusion/ neg)**	***Domain_L***
Yes	446	274	446	274	0.673716012	0.006724752	100.1845
No	216	40471	662	40745	0.326283988	0.993275248	0.328493
Sum	***662***	***40745***					

2.3 Bayesian Networks for PPI Prediction

Bayesian networks have incredible power to offer assistance in a wide range of endeavors. They support the use of probabilistic inference to update and revise belief values. Bayesian networks readily permit qualitative inferences without the computational inefficiencies of traditional joint probability determinations [5]. Furthermore, as integrating multiple data source has been emphasized in recent bioinformatics, Bayesian networks are quite suitable for the task of combining various features from heterogeneous data sources.

Bayesian networks are a representation of the joint probability distribution among multiple variables (which could be datasets or information sources). Denote by 'positive' a pair of proteins that are in the same complex. Given the number of positives among the total number of protein pairs, the 'prior' odds of finding a positive are:

$$O_{prior} = \frac{P(pos)}{P(neg)} = \frac{P(pos)}{1 - P(pos)}$$

In contrast, the 'posterior' odds are the odds of finding a positive after we consider N data sets with feature values $f_1 \ldots f_N$:

$$O_{post} = \frac{P(pos | f_1 \ldots f_N)}{P(neg | f_1 \ldots f_N)}$$

We estimated the likelihood ratio L for each feature value:

$$L(f_1 \ldots f_N) = \frac{P(f_1 \ldots f_N | pos)}{P(f_1 \ldots f_N | neg)}$$

In this paper, we assumed that three genomic features, MIPS functional catalog, GO biological process and domain fusion, are independent, and then we applied naïve Bayesian networks to infer PPI from these features. In that case, L can be simplified to:

$$L(f_1 \ldots f_N) = \prod_{i=1}^{N} L(f_i) = \prod_{i=1}^{N} \frac{P(f_i | pos)}{P(f_i | neg)}$$

The parameters of the naïve Bayesian network are shown in Table 1. Columns "pos" and "neg" give the overlap of protein pairs in sampling dataset with the 8,250 positives and the 2,708,746 negatives. Columns "sum (pos)" and "sum(neg)" show the number of positives/negatives among the protein pairs with likelihood ratio greater than or equal to *L*. *P(feature value/positive)* is the probability of a positive having a corresponding feature value and *P(feature value/negative)* is the probability of a negative having a corresponding feature value.

From above equation, the likelihood ratios *MIPS_L*, *GO_L*, and *Domain_L* showed in the last columns are the likelihood ratios for respective feature values extracted from databases MIPS, GO, and Domain fusion. We obtained very high *MIPS_L, GO_L* and *GO_L*, especially *Domain_L* (*Domain_L* = 100.1845). The likelihood ratio *Domain_L* of domain fusion feature is the highest ratio *L* among the ratios of various features extracted from other data sources used in [8], [9]. This means that domain fusion is the helpful genomic feature to predict reliable protein interactions.

Likelihood ratios of all genomic features are multiplied to estimate the final one, likelihood ratio *L*, as the representation of integrating of genomic features in naïve Bayesian networks and then increase the accuracy of protein interactions:

$$L = MIPS\_L * GO\_L * Domain\_L$$

In this paper, we have not applied these likelihood ratios to the whole yeast interactome (about 18 millions pairs) yet; we only did experiments with a dataset of about 7 millions pairs from three genomic data sources. This dataset is rich information because in theory all protein pairs in this dataset are likely to interact.

In Section 3, the experimental results show that Bayesian networks are suitable for integrating many different features. The evaluation with different aspects demonstrates that domain fusion is the appropriate feature to reliably predict PPI and open some further studies.

3. Experimental Results and Evaluation

We chose different thresholds of *L* to predict PPI. Doing the experiments with the whole data set (7 millions protein pairs), we obtained some good results (as shown in Figure 2). In Figure 2, we can see that when the threshold *L* increases, the

number of protein will decrease. But we think that $L \geq 25$ is a suitable threshold for determining a protein-protein interaction [13]. With this threshold, we predicted 70,601 protein interactions. When $L \geq 1$ (means that the probability of a protein pair as positive is higher than as a negative), there are 443,896 protein pairs that satisfy this threshold. This means that the proposed method is effective to predict protein interactions. The big number of predicted protein interactions encouraged us to follow this approach.

In Jansen *et al.*'s work [8], they did several experimental evaluations such as comparison of Bayesian networks with voting, cross-validation.

In the different ways, we carried out various experiments and other evaluation to validate the reliability of our work. First, we evaluated the predicted PPI by comparing them with other well-known PPI data sets such as CYGD (Comprehensive Yeast Genome Database) [24] and DIP (Database of Interacting Proteins) [19, 25]. If in the predicted PPI data set, there are many overlapping PPI with DIP and CYGD, we could rely on the predicted results. The results are showed in Figure 3. A large amount of the predicted PPI overlaps in DIP database and CYGD database.

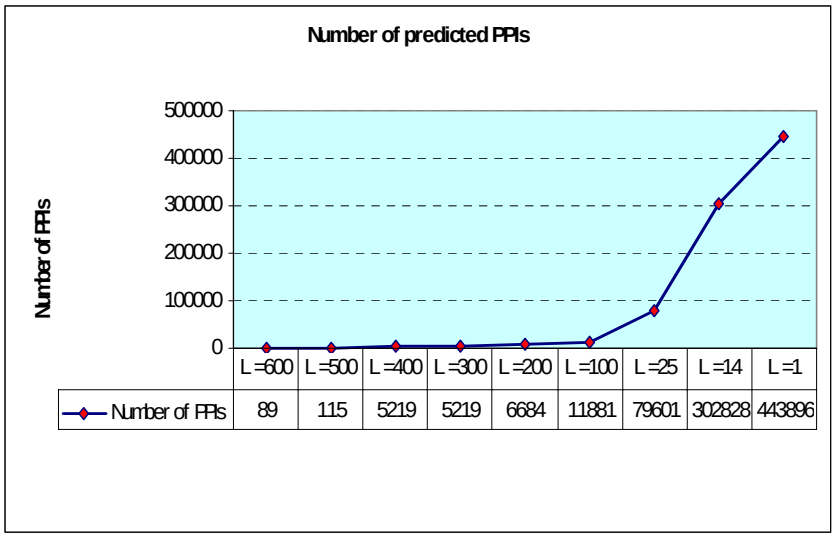


Figure 2. Number of predicted PPI with various thresholds *L*.

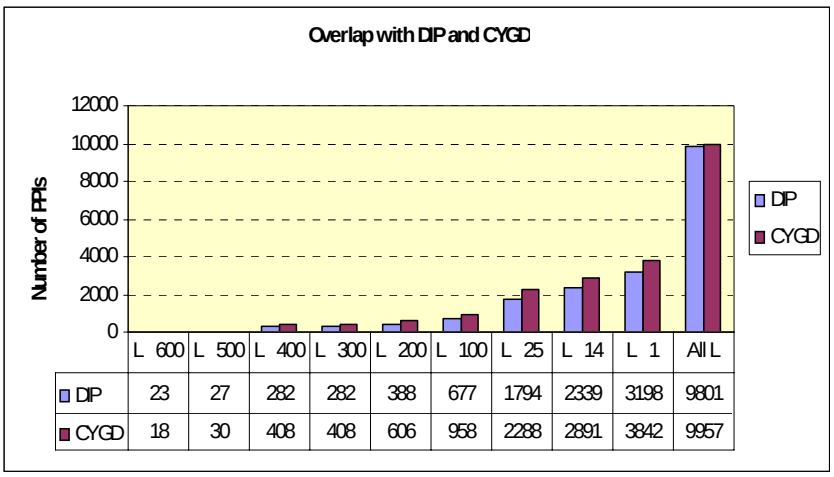
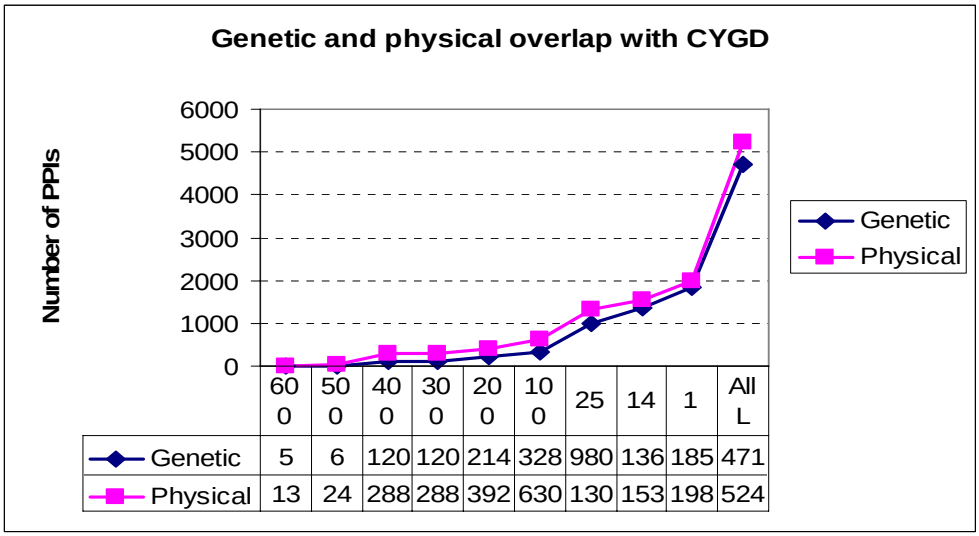


Figure 3. Comparison with overlapped PPI in databases DIP and CYGD .

Second, owing to DIP database, we investigated the characteristic of predicted PPI, whether they are physical or genetic PPI. In biology, PPI are divided into two types: physical PPI and genetic PPI [4]. Physical PPI, which are considered as the direct ones, play a very important role in forming stable complexes. These complexes perform biological roles together and can last a long time in cell. Because protein domains are the basic parts of proteins then we expected with domain fusion features we can better predict physical PPI. In some situations, stable physical interactions are thought to be more reliable than transient genetic ones. Figure 4 shows the number of physical and genetic PPI corresponding to threshold L. With $L \geq 500$, all of PPI are physical. For every thresholds of L, the number of physical PPI is always higher than the number of genetic PPI. These results mean that our method promisingly predict physical PPI.

Third, we verified the reliability of the putative PPI by looking up their domain-domain interactions in InterDom database [11]. Like proteins, protein domains also interact together, and they are considered as the stable channels behind protein-protein interactions. Recently, many studies take domain-domain interactions (DDI) into account to predict PPI. From that point of view, we tried to validate the predicted PPI through DDI. The number of PPI having at least one DDI is given in Figure 5. From more than 30,000 DDI extracted from InterDom database, there are lot of predicted PPI verified to have at least one DDI.

To build up complexes of proteins or protein pathways, we detected combination groups of interactions, in which protein A interacts with protein B and protein B interacts with protein C. Such kinds of groups are called triplets. With $L \geq 25$ then, there are a huge number of 1,851,579 triplets. We may think about the way to build the cellular pathways and protein complexes from these triplets. Actually, the cellular pathways in cells such as metabolism pathways or signal transduction pathways are much more complicated, but these triplets could be cores of the cellular pathways. Spreading groups with four, five or more protein interactions, we can construct the complex networks of protein interactions and cellular pathways.



	600	500	400	300	200	100	25	14	1	All L
Genetic	5	6	120	120	214	328	980	136	185	471
Physical	13	24	288	288	392	630	130	153	198	524

Figure 4. Genetic and physical protein-protein interactions overlapped with those in CYGD database.

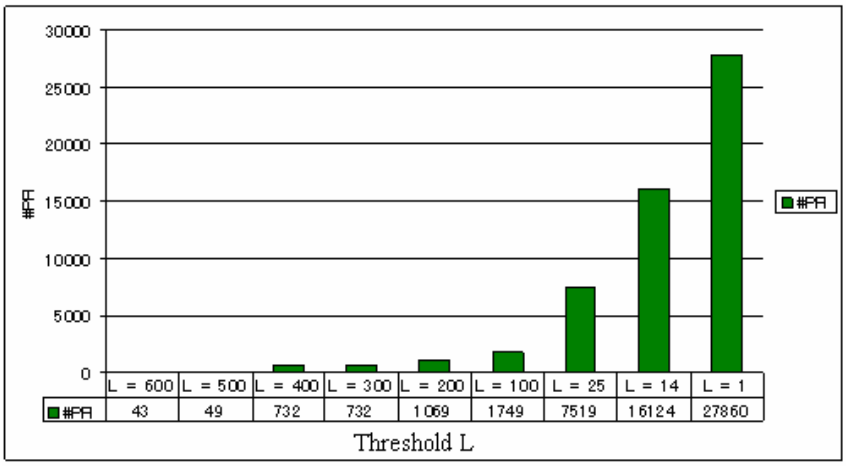


Figure 5. Number of PPI having DDI in InterDom database.

4. Discussion and Conclusion

In this work, we applied Bayesian networks to predict novel and reliable protein interactions in yeast. We chose three predictive genomic features in particular domain fusion to discover protein-protein interactions. Among various genomic features, domain fusion stands out as the promising feature to predict protein-protein interactions. The efficiency of our method is not only the huge number of predicted protein-protein interactions but also the reliability of them. And these putative proteins are carefully validated by various ways.

The result will be better when the data missing problem can be solved. The missing data problem is that we have not had an adequate domain fusion database. But these problems can be resolved as genomic data is more available. The method presented in this paper can be also applied for various organisms and alternative genomic features. In future, we would like to build up the protein complexes, and then cellular pathways. To improve the method, other genomic/proteomic features need combing to have more comprehensive view of protein-protein interactions.

References

[1] Arnau, V., Mars, S. and Martin, I. (2005). "Iterative Cluster Analysis of Protein Interaction Data", *Bioinformatics* 21(3), 364-378.

[2] Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21(suppl1), i38–46.

[3] Ben-Hur, A. and Noble, W. S. (2006). Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, 7(suppl 1).

[4] Chen, Y. and Xu, D. (2003). "Computational Analyses of High-Throughput Protein-Protein Interaction Data", *Current Protein and Peptide Science*, 4(3),159-180.

[5] Cooper, G.F., and Herskovits, E., "A Bayesian Method for the Induction of Probabilistic Networks from Data," *Machine Learning*, 9: 309-347, 1992.

[6] Deane, C.M., Salwinski, L., Xenarios, I., Eisenberg, D., "Protein interactions: two methods for assessment of the reliability of high throughput observations", *Mol Cell Proteomics,* 1(5):349-56, 2002.

[7] Friedrich, T., Pils, B., Dandekar, T., Schultz, J. and Muller, T. (2006). "Modelling interaction sites in protein domains with interaction profile hidden Markov models", *Bioinformatics* 22(23), 2851-2857.

[8] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., Gerstein, M., "Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data", *Science*, 302: 449 – 453, 2003.

[9] Lu, L.J., Xia, Y., Paccanaro, A., Yu, H., Gerstein M., Assesing the limits of genomic

data integration for predicting protein networks, *Genome Research*, 15:945 - 953, 2005.

[10] Marcotte, E., Pellegrini, M., Ho-leung, Rice, D., Yeates, T., Eisenberg, D., "Detecting Protein Function and Protein – Protein Interactions form Genome Sequence", *Science*, 285: 751 – 753, 1999.

[11] Ng, S., Zhang, Z., Tan, S. and Lin, K. (2003). "InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes", *Nucleic Acids Res* 31(1), 251-254.

[12] Oyama, T., Kitano, K., Satou, K., Ito, T., "Extraction of knowledge on PPI by association rule discovery", *Bioinformatics*, 18: 705 – 714, 2002.

[13] Pellegrini, M., Marcotte, E.M., Thompson, M.J. , Isenberg, D., Yeates, T.O, "Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles", *Proc. Natl. Acad. Sci. USA*, 96: 4285–4288, 1999.

[14] Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. (2005). Probabilistic model of the human protein-protein interaction network. *Nat Biotech,* 23(8), 951–959.

[15] Salwinski, L., Eisenbergy, D., "Computational methods of analysis of protein–protein interactions", *Curr Opin Struct Biol*, 13(3):377-82, 2003.

[16] Schölkopf, B., Tsuda, K., and Vert, J.P, (2004). Kernel Methods in Computational Biology, The MIT Press.

[17] See-kiong, Ng., Tan, S., "Discovering Protein-Protein Interactions", J*ournal of Bioinformatics and Computational Biology*, Vol. 1. No. 4, Imperial College Press, 2004.

[18] Truong, K., Ikura, M., "Domain fusion analysis by applying relational algebra to protein sequence and domain databases", *BMC Bioinformatics*, 2003.

[19] Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., Eisenberg, D., "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions", *Nucleic Acids Res*, 30(1):303-5, 2002.

[20] Yuan, B., "Whole Genome Analysis for Protein-Protein Interactions in Yeast", *Genome Research*, 12:37– 46, 2002.

[21] http://www.geneontology.org/.

[22] http://mips.gsf.de/.

[23] http://calcium.uhnres.utoronto.ca/pi.

[24] http://mips.gsf.de/genre/proj/yeast/.

[25] http://dip.doe-mbi.ucla.edu/.