Inferring Protein-Protein Interactions Using a Hybrid Genetic Algorithm/Support Vector Machine Method

Bing Wang[1,2,*], Peng Chen[3], Jun Zhang[2], Guangxin Zhao[1] and Xiang Zhang[2]

[1]School of Electrical & Information, Anhui University of Technology, Ma'anshan, Anhui, 243002, China; [2]Department of Chemistry, University of Louisville, Louisville, KY, 40202, USA; [3]School of Computer Engineering, Nanyang Technological University, Singapore, 639798, Singapore

Abstract: Identifying protein-protein interaction is crucial for understanding the biological systems and processes, as well as mutant design. This paper proposes a novel hybrid Genetic Algorithm/Support Vector Machine (GA/SVM) method to predict the interactions between proteins intermediated by the protein-domain relations. A protein domain is a structural and/or functional unit of the protein. Every protein can be characterized by a distinct domain or a sequential combination of multiple domains. In our method, the protein was first represented by its domains where the effects of domain duplication were also considered. Transformation of the domain composition was taken to simulate the combination of different domains using genetic algorithm (GA). The optimal transformation was discovered using a predictor constructed by a support vector machines (SVM) method. Compared with random predictor, the prediction performance of our method is more effective and efficient with 0.85 sensitivity, 0.90 specificity and 0.88 accuracy.

Keywords: Protein-protein interaction, protein-domain relations, genetic algorithm, support vector machine, domain composition, composition transformation.

1. INTRODUCTION

The study of protein-protein interactions is one of the most important topics in molecular biology. The interactions between proteins play a critical role in the live biological cells by controlling such as regulation of metabolic and signaling pathways, immunologic recognition, DNA replication and gene translation, and protein synthesis [1]. Several experimental methods have been developed to analyze protein-protein interactions, including yeast two-hybrid assay [2-5], protein chips [6], and mass spectrometry of purified protein complexes [7, 8]. However, each of these techniques is tedious, time-consuming and labor-intensive [9], and suffers from high rates of both false positive and false negative predictions [10, 11]. Therefore, it is becoming important for researchers to seek computational approaches, which are much faster and less expensive than experimental analyses, to facilitate the development of interactomics by predicting protein interactions.

A number of computational techniques have been suggested for the prediction of potential protein interactions and theirs interaction sites [12-17]. Some of them focus on the functional relationships between proteins, such as the Gene fusion (Rosetta stone) method [9, 18] and the phylogenetic profiles [19, 20]. The others emphasize particularly on structural interactions and related information, such as probabilistic model methods [21, 22], evolutionary important residues clustering [23], domain pair profile method [24, 25].

Recently, methods based on the relationships between proteins and domains had been developed to study protein-protein interactions. These domain-based approaches are motivated by the fact that protein-protein interactions are mediated by the physical interactions between their domains. Ng *et al*. [26] developed an automated interacting domain discovery system, referred to as InterDom, based on an integrative approach. Kim *et al*. [27] and Han *et al*. [28] presented several statistical methods, which are similar to the association method described by Sprinzak and Margalit [29]. The precision of these approaches suffers from the ignorance of experimental errors in the proposed models. Deng *et al*. [30] applied the Maximum Likelihood Estimation (MLE) method to infer the domain-domain interactions from protein-protein interactions, which had been shown to be of robustness in dealing with various experimental errors. Furthermore, Gomez *et al*. constructed an attraction-repulsion model associated with Pfam domains [22].

Although these studies showed that the protein-domain relationships involved protein interaction information and some progresses had been made toward protein interaction prediction, most of them assumed that two proteins interact if and only if at least one pair of domains from the two proteins interacts and domain-domain interactions are independent with each other. That is to say current methods are based on that single-domain pair is the basic unit of protein interactions. However, it is apparent that multiple domains participate in physical interaction in multiple complex structures. It is reasonable that the possible domain combinations will contribute to the protein interactions.

In this paper, we propose a novel hybrid Genetic Algorithm/Support Vector Machine (GA/SVM) method to tackle the prediction of protein interactions based on protein do-

*Address correspondence to this author at the School of Electrical & Information, Anhui University of Technology, Ma'anshan, Anhui, 243002, China; Tel: +86 555 2311541; Fax. +86 555 2311540;
E-mail: wangbing@ustc.edu

0929-8665/10 $55.00+.00

main composition. We first characterized a protein by its domains. Here, we not only considered the types of domains, but also the number of domains. For a protein pair, a feature vector was constructed by concatenation of each feature of the two proteins. We also considered the prediction of protein-protein interaction as a two-class classification problem: interaction and non-interaction. A transformation of domain composition was adopted to include the effects of multiple domains and to reduce the dimensions of input vector. Specifically, a GA algorithm was used to discover an optimal transformation to enhance the prediction performance of our proposed SVM predictor. Experimental results have shown that the domain composition indeed can be used to infer protein-protein interactions and that the optimized transformation results in a significant improvement in prediction performance.

2. EXPERIMENTS AND ALGORITHM

2.1. Dataset Preparation

Protein-protein interaction data was collected from *Saccharomyces cerevisiae* core subset of DIP database [31]. This dataset was validated by two methods described by Deane and colleagues [32]. The first is to use the expression profile reliability index to estimate the biologically relevant fraction of protein interactions by comparing the RNA expression profiles of the proteins with expression profiles of known interacting and non-interacting pairs of proteins. The second is to use the paralogous verification method to test the reliability of a putative interaction pair by examining whether there is a known paralog that also interacts with its partner protein. The DIP database includes 5951 protein-protein interactions in yeast organism.

Domain information can be obtained from Pfam database [33]. Pfam database contains a large collection of multiple sequence alignments and profile hidden Markov models (HMM) covering the majority of protein domains. There are 1943 Pfam domains in the current version of Pfam (v19.0). By mapping these Pfam domains to the core subset of DIP database, we got 3611 protein pairs that each protein pair has at least one Pfam domain in both interacting proteins. After excluded the domains not found in the all protein pairs, only 1874 Pfam domains were left and used in this study. Therefore, the positive interacting protein dataset used here includes 3611 protein pairs, whose correlative domain's number is 1874.

Due to non-interacting protein information is not available currently, a hypothetical non-interacting protein dataset was generated based on subcellular localization information. It consisted of protein pairs that do not co-localize together. The subcellular localization source was extracted from Munich Information Center for Protein Sequences (MIPS) [34] and only the four main types of localization were considered in this study – cytoplasm, nucleus, mitochondria and endoplasmic reticulum. The yeast proteins used in the positive dataset were assigned with the four types of localization information and those with multiple localizations were removed to minimize the introduction of potential noise during the training process. Four sets of proteins with respect to the four types of localization were generated, and proteins from each set were subsequently paired with proteins from a different localization. In the experiment reported here, the whole dataset was formed by the positive subset and the negative subset at about 1:1 ratio. Although there are enormous amount of possible negative pairing, 5000 protein pairs were randomly selected in this work to start with. After removing duplication and performing exclusion analysis of the whole DIP yeast interacting proteins, 4660 protein pairs were used as the hypothetical non-interacting dataset.

2.2. Feature Representation

In this work, the protein-protein interaction prediction problem was formulated as a two-class classification problem: each protein pair is a sample belonging to either 'interaction' class (the two proteins interact with each other) or 'non-interaction' class (the two proteins do not interact with each other). In our application, a protein was characterized by the domains existing in each protein. The feature vector of each protein therefore can be formulated as

$$p = [d_1, d_2, ..., d_i, ..., d_n] \tag{1}$$

where each feature corresponds to a kind of domain existing in the protein, the value of d_i is the number of this type of domain, and $d_i = 0$ otherwise. The domain composition of each protein was extracted from the Pfam database by using InterProScan [35]. The effects of domain duplication can be taken into account by using this formula to construct the feature vector.

Each protein pair is represented by the domains of two proteins. The full feature vector for a particular protein pair was constructed by concatenation of each feature of the two proteins. This can be written as $x = p_1 \oplus p_2$, the length of full feature vector therefore is 2n, where $n = 1874$ is the number of all kinds of protein domains we used here Fig. (**1**).

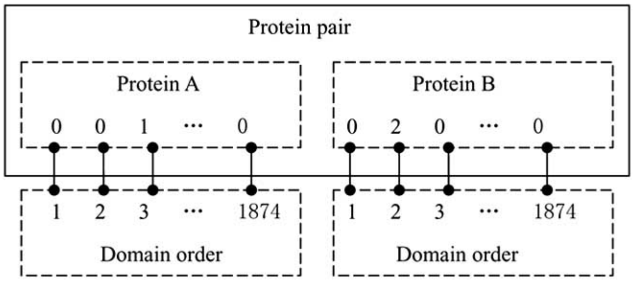

Figure 1. Representation of protein pair using domain composition. Here, protein is consisted of protein A and protein B, where protein A has one domain 3 and protein B has two domain 2.

Most domain-based protein-protein prediction methods infer domain-domain interacting information from protein-protein interaction and then try to predict protein interactions based on the inferred domain-domain interacting information. However, those approaches only considered the interactions of single domain pairs and assigned an interacting probability for a specific protein pair based on the biggest probability or a combination of probability values of the domain pairs within the protein pair. Therefore, the association of different domains no matter what is same domain type or not will be ignored in the protein pair which is consisted of multiple domains. In our proposed protein representation, all the domains within the protein are taken into account will allow us to consider the effect of the combination of domains.

2.3. Hybrid GA/SVM Classifier

It is known that the assumption that domain-domain interactions are independent for each other is not biologically reasonable because two or more domains may cooperatively interact with another domain. In addition, interaction domains are often used repeatedly by many multi-domain proteins, and the evolution and spread of domains through different protein families exists as a result of gene duplication. Our original representation of protein gives us a chance to deal with those problems, but it does not take the similarity and evolution of domains into account and it therefore, may not provide a satisfying classification rate between the interacting protein pairs and the non-interacting pairs. Another issue about the original protein pair's representation is that the dimensionality of the feature vectors is very high. As structural, functional and evolutionary units of proteins, only generally several domains exist in the multiple-domain protein. There is therefore small number of non-zero features in the original vector of protein pairs. Apparently, it is necessary to reduce the dimension of protein representation. In this work, we adopted a hybrid GA/SVM classifier to address those problems and predict interacting protein pairs.

In the construction of the hybrid GA/SVM classifier, we first transformed original domain composition to enhance classification rate and to reduce the dimensionality of the domain features. Meanwhile, the transformation of original domain composition will simulate the different combination possibility among the domains. Then, the GA was used to create different types of transformation while the SVM was adopted to evaluate the effectiveness of corresponding transformation.

a. Transformation of Domain Composition

In this work, we took the value of each feature as a histogram and the height of each bin in the histogram representation denotes the corresponding number of domains. The objective of current approach is to transform the original protein representation into a relative lower dimensional vector to classify a protein pair into interacting set or non-interacting set. This operation is similar to constructing a mapping $f: R^N \to R^M$ (M<N) and the reduction of dimensionality can be realized by selective merging the value of original features.

It is apparent that a number of different transformations may exist and can result in different classification performance. Our basic assumption is that there should exist a set of suitable transformations by which better classification results can be obtained. This assumption thus applies to the protein pair representation, which is basically a histogram of domain composition. In general, the resulting histograms may not be adequate for characterizing the similarity and/or association relationships among domains. As a result, we need to transform these histograms in such a way to capture and to magnify these similarity and/or association relationships more closely, while retaining the computational advantages of our chosen representation. Our objective is therefore to pursuit the best transformation by which we can achieve the best classification performance.

b. Selection of the Optimized Transformation Using GA/-SVM

Genetic algorithm (GA) is a randomized search and optimization technique guided by the principles of evolution and natural genetics. In order to find out the optimal solution of a problem, a GA starts from a set of assumed solutions (chromosomes) and evolves different but better sets (of solutions) over a sequence of iterations. In each generation (iteration) the objective function (fitness measuring criterion) determines the suitability of each solution and, based on these values, some genetic operations (selection/reproduction, crossover and mutation) are operated to produce the next generation. In general, GAs can rapidly locate good solutions, even for difficult search spaces [36].

The process of transforming is illustrated in Fig. (**2**). We encoded the chromosome into a character string whose length equals to the size of the population. A ternary alphabet $\Lambda = \{a,b,c\}$ was adopted for the strings. In the transformation process of domain composition, if the consecutive characters in the chromosome are identical, the corresponding positions in the original vector were merged and their values were combined to a numeral.

In the GA/SVM method, SVM was used to evaluate each candidate transformation. According to the transformations, we recombined the domain compositions and used it as input vector of the SVM predictor. The dataset was randomly partitioned into two subsets with almost same size. Consequently, two SVM classifiers were trained using each of them and tested on its complement subset. The fitness function was then defined as the average classification rate of these two SVM classifiers.

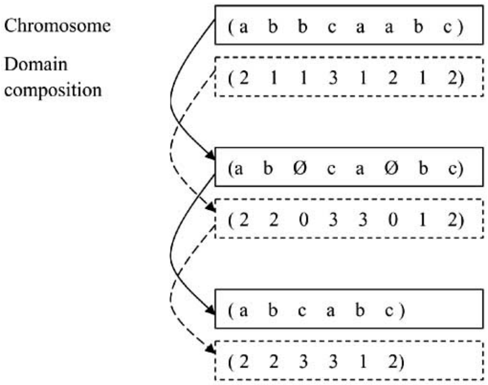

Figure 2. The transformation process of domain composition. 1.Merge identical characters in chromosome; 2. Delete empty position in chromosome; accordingly, 1'.Combine the value in domain composition in corresponding position; 2'.Delete zero value in domain composition.

3. RESULTS

3.1. Evaluation Criteria

We used three measures to evaluate our method of predicting protein-protein interactions: sensitivity, specificity and accuracy. The sensitivity is defined as the ratio of the number of matched interacting protein pairs over the total number of the positive samples in the observed set. The specificity is generally defined as the ratio of the number of matched non-interacting protein pairs between the predicted set and the actual set over the total number of negative samples. Let TP be the number of true positives, i.e. protein pairs predicted to be interacting pairs that actually are interaction pairs, and FP be the number of false positives, i.e. protein pairs predicted to be interacting protein pairs that are in fact not interaction pairs. In addition, let TN be the number of true negatives, and FN the number of false negatives. Then the evaluation measures can be computed as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{4}$$

3.2. Prediction Performance

In our genetic algorithm, the population is set to contain 50 individuals and the selection procedure uses the roulette wheel algorithm. The possibility of crossover is set to be 0.8 and the crossover type is selected as single point, while the mutation probability is 0.01. The termination criterion is based on fitness value that best score does not change over 50 generations.

In the process of genetic optimization, transformed domain compositions can be obtained by removing and merging some bins in the histogram representation of protein pairs, which dramatically reduce the computational complexity while still maintain the domain composition information. Each individual in the population in GA corresponds to a possible combination of the original domains. For instance, the optimized domain composition is found by our hybrid GA/SVM method by which best protein-protein prediction performance can be achieved. In this transformed domain composition, the dimensionalities of input vector are reduced to 1151 after the optimization process, which is only 64.5% of original dimensionality. This significant reduction of data dimensionality decreases computational complexity dramatically and makes our method much faster in predicting new protein-protein interactions.

The general performances of our method are in Table **1**. It can be seen that our domain composition based predictor generated higher values (> 0.8) in all of the three performance measures: sensitivity, specificity and accuracy. Those results indicated that the false positives and false negatives are very small, which is very significant for protein interaction prediction because current experimental techniques involved high false positives and false negative problems. To evaluate the effectiveness of the domain composition transformation in our proposed method, we constructed a random predictor in which we shuffled the predictions and assigned them to the protein pairs in the test set randomly. This process is important for it enables us to infer the significance of our results. According to the performance information provided in Table **1**, the values of the three performance measures generated by the random predictor are around 0.5. The higher value of correlation coefficient (> 0.75) demonstrated that our GA/SVM predictor significantly outperformed the random predictor. The corresponding performance achieved by our proposed approach are higher than that of random predictor apparently indicates that domain-based protein pair representation indeed contains protein function information and it can be helpful for the identification of protein-protein interactions.

For further estimate the contribution of the transformation of domain composition, we compare the prediction results of hybrid GA/SVM method with that of SVM classfier which use the original domain composition as input vector. From Table **1**, it can be seen that the transformation of domain composition can enhance prediction performance significantly. Selecting the transformation of domain composition using GA leads to an impressive improvement in performance compared to the original domain representation: at least 2.2% increase in *sensitivity*, 10.8% increases in *specificity*, 0.065 in accuracy, and 12% in *correlation coefficient*. The result shows that the transformation of domain composition can represent protein-protein interaction information more effectively for it consider the similarity and association of different domains, which is important in protein-protein interaction, especially for multiple-domain proteins.

CONCLUSION

In this paper, we proposed a new method for the prediction of protein-protein interactions from the protein-domain relationships. Unlike most of previous approaches which tackled the same problem based on an assumption that domain interactions are independent with each other, our

Table 1. The General Performance of Prediction

	Sensitivity	Specificity	Accuracy	Correlation coefficient
SVM_randmom[a]	0.5209	0.5208	0.5209	0.0058
SVM[b]	0.8319	0.7959	0.8139	0.6382
GA/SVM[c]	0.8543	0.9045	0.8794	0.7589

a) SVM_random is the predictor whose prediction results are shuffled randomly;
b) SVM is the predictor whose attributes are original domain composition information;
c) GA/SVM is the predictor whose attributes are transformed domain composition information using GA algorithm.

method can consider the multiple domain effects. Specifically, we treated the prediction of protein-protein interactions as a two-class classification problem. We also adopted an optimized transformation of domain composition, which was selected using GA algorithm to construct the SVM predictor. The experimental results demonstrated that the prediction performance was significantly enhanced by using of our hybrid GA/SVM method. The better experimental results indicate that our proposed GA/SVM predictor could capture the difference between the interacting protein pairs with the non-interacting pairs. Furthermore, the performance of our method can be further improved when the domain information is further and more reliably annotated.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation of China (No. 60803107).

REFERENCES

[1] Alberts, B. *Molecular biology of the cell*. 4th ed.; Garland Science: New York, **2002**.

[2] Ito, T.; Tashiro, K.; Muta, S.; Ozawa, R.; Chiba, T.; Nishizawa, M.; Yamamoto, K.; Kuhara, S.; Sakaki, Y. 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. *Proc. Natl. Acad. Sci*., **2000**, *97*(3), 1143-7.

[3] Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M.; Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci*., **2001**, *98*(8), 4569-74.

[4] Ito, T.; Chiba, T.; Yoshida, M. Exploring the protein interactome using comprehensive two-hybrid projects. *Trends. Biotechnol*., **2001**, *19*(10 Suppl), S23-7.

[5] Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T. A.; Judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S.; Rothberg, J. M. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **2000**, *403*(6770), 623-7.

[6] Zhu, H.; Bilgin, M.; Bangham, R.; Hall, D.; Casamayor, A.; Bertone, P.; Lan, N.; Jansen, R.; Bidlingmaier, S.; Houfek, T.; Mitchell, T.; Miller, P.; Dean, R. A.; Gerstein, M.; Snyder, M. Global analysis of protein activities using proteome chips. *Science*, **2001**, *293*(5537), 2101-5.

[7] Gavin, A. C.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A. M.; Cruciat, C. M.; Remor, M.; Hofert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M. A.; Copley, R. R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B.; Neubauer, G.; Superti-Furga, G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **2002**, *415*(6868), 141-7.

[8] Ho, Y.; Gruhler, A.; Heilbut, A.; Bader, G. D.; Moore, L.; Adams, S. L.; Millar, A.; Taylor, P.; Bennett, K.; Boutilier, K.; Yang, L.; Wolting, C.; Donaldson, I.; Schandorff, S.; Shewnarane, J.; Vo, M.; Taggart, J.; Goudreault, M.; Muskat, B.; Alfarano, C.; Dewar, D.; Lin, Z.; Michalickova, K.; Willems, A. R.; Sassi, H.; Nielsen, P. A.; Rasmussen, K. J.; Andersen, J. R.; Johansen, L. E.; Hansen, L. H.; Jespersen, H.; Podtelejnikov, A.; Nielsen, E.; Crawford, J.; Poulsen, V.; Sorensen, B. D.; Matthiesen, J.; Hendrickson, R. C.; Gleeson, F.; Pawson, T.; Moran, M. F.; Durocher, D.; Mann, M.; Hogue, C. W.; Figeys, D.; Tyers, M. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **2002**, *415* (6868), 180-3.

[9] Enright, A. J.; Iliopoulos, I.; Kyrpides, N. C.; Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **1999**, *402*(6757), 86-90.

[10] Legrain, P.; Selig, L. Genome-wide protein interaction maps using two-hybrid systems. *FEBS Lett.*, **2000**, *480* (1), 32-6.

[11] Edwards, A. M.; Kus, B.; Jansen, R.; Greenbaum, D.; Greenblatt, J.; Gerstein, M. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends. Genet*., **2002**, *18*(10), 529-36.

[12] Shi, M. G.; Huang, D. S.; Li, X. L. A protein interaction network analysis for yeast integral membrane protein. *Protein Pept. Lett*., **2008**, *15*(7), 692-9.

[13] Wang, B.; Chen, P.; Huang, D. S.; Li, J. J.; Lok, T. M.; Lyu, M. R. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett*., **2006**, *580*(2), 380-4.

[14] Wang, B.; Wong, H. S.; Chen, P.; Wang, H. Q.; Huang, D. S. Predicting Protein-Protein Interaction Sites Using Radial Basis Function Neural Networks. In *International Joint Conference on Neural Networks*, **2006**; pp. 2325-2330.

[15] Wang, B.; Wong, H. S.; Huang, D. S. Inferring protein-protein interacting sites using residue conservation and evolutionary information. *Protein Pept. Lett.*, **2006**, *13*(10), 999-1005.

[16] Chen, P.; Han, K.; Li, X.; Huang, D. S. Predicting key long-range interaction sites by B-factors. *Protein Pept. Lett.*, **2008**, *15*(5), 478-83.

[17] Xia, J. F.; Han, K.; Huang, D. S. Sequence-Based Prediction of Protein-Protein Interactions by Means of Rotation Forest and Autocorrelation Descriptor. *Protein Pept. Lett*., **2010**, *17*(1), 137-145.

[18] Marcotte, E. M.; Pellegrini, M.; Ng, H. L.; Rice, D. W.; Yeates, T. O.; Eisenberg, D. Detecting protein function and protein-protein interactions from genome sequences. *Science*, **1999**, *285*(5428), 751-3.

[19] Marcotte, E. M.; Pellegrini, M.; Thompson, M. J.; Yeates, T. O.; Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature*, **1999**, *402*(6757), 83-6.

[20] Pellegrini, M.; Marcotte, E. M.; Thompson, M. J.; Eisenberg, D.; Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci*., **1999**, *96*(8), 4285-8.

[21] Gomez, S. M.; Lo, S. H.; Rzhetsky, A. Probabilistic prediction of unknown metabolic and signal-transduction networks. *Genetics*, **2001**, *159*(3), 1291-8.

[22] Gomez, S. M.; Noble, W. S.; Rzhetsky, A. Learning to predict protein-protein interactions from protein sequences. *Bioinformatics*, **2003**, *19*(15), 1875-81.

[23] Lichtarge, O.; Yao, H.; Kristensen, D. M.; Madabushi, S.; Mihalek, I. Accurate and scalable identification of functional sites by evolutionary tracing. *J. Struct. Funct. Genomics*, **2003**, *4*(2-3), 159-66.

[24] Wojcik, J.; Schachter, V. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, **2001**, *17 (Suppl 1)*, S296-305.

[25] Rain, J. C.; Selig, L.; De Reuse, H.; Battaglia, V.; Reverdy, C.; Simon, S.; Lenzen, G.; Petel, F.; Wojcik, J.; Schachter, V.; Chemama, Y.; Labigne, A.; Legrain, P. The protein-protein interaction map of Helicobacter pylori. *Nature*, **2001**, *409*(6817), 211-5.

[26] Ng, S. K.; Zhang, Z.; Tan, S. H. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, **2003**, *19*(8), 923-9.

[27] Kim, W. K.; Park, J.; Suh, J. K. Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. *Genome Inform*., **2002**, *13*, 42-50.

[28] Han, D.; Kim, H. S.; Seo, J.; Jang, W. A domain combination based probabilistic framework for protein-protein interaction prediction. *Genome Inform*., **2003**, *14*, 250-9.

[29] Sprinzak, E.; Margalit, H. Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol*., **2001**, *311*(4), 681-92.

[30] Deng, M.; Mehta, S.; Sun, F.; Chen, T. Inferring domain-domain interactions from protein-protein interactions. *Genome Res*., **2002**, *12*(10), 1540-8.

[31] Salwinski, L.; Miller, C. S.; Smith, A. J.; Pettit, F. K.; Bowie, J. U.; Eisenberg, D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids. Res*., **2004**, *32*(Database issue), D449-51.

[32] 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*, **2002**, *1*(5), 349-56.

[33] Bateman, A.; Coin, L.; Durbin, R.; Finn, R. D.; Hollich, V.; Griffiths-Jones, S.; Khanna, A.; Marshall, M.; Moxon, S.; Sonnhammer, E. L.; Studholme, D. J.; Yeats, C.; Eddy, S. R. The Pfam protein families database. *Nucleic Acids Res*., **2004**, *32*(Database issue), D138-41.

[34] Mewes, H. W.; Dietmann, S.; Frishman, D.; Gregory, R.; Mannhaupt, G.; Mayer, K. F.; Munsterkotter, M.; Ruepp, A.; Spannagl, M.; Stumpflen, V.; Rattei, T. MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res*., **2008**, *36*(Database issue), D196-201.

[35] Quevillon, E.; Silventoinen, V.; Pillai, S.; Harte, N.; Mulder, N.; Apweiler, R.; Lopez, R. InterProScan: protein domains identifier. *Nucleic Acids Res*., **2005**, *33*(Web Server issue), W116-20.

[36] Zhao, X. M.; Huang, D. S.; Cheung, Y. M. A novel hybrid GA/RBFNN technique for protein classification. *Protein Pept. Lett.*, **2005**, *12*(4), 383-386.

Received: April 29, 2010 **Revised: May 05, 2010** **Accepted: June 14, 2010**