18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Fuzzy systems modeling for protein-protein interaction prediction in *Saccharomyces cerevisie*

Abu Bakar, Sakhinah, Javid Taheri, and Albert Y. Zomaya

School of Information Technologies, Faculty of Engineering, The University of Sydney, New South Wales 2006, Australia.
Email: sabu0366@it.usyd.edu.au

Abstract: Most of the biological functions are mediated by protein-protein interactions in the organism. If one of these interactions behaves improperly, it may lead to a disease. Therefore, the study of protein-protein interactions is very important to improve our understanding of diseases and can provide the basis for new therapeutic approaches. Although, there are no concrete properties in predicting protein-protein interactions, it is known from experimentally determined protein-protein interactions that interacting proteins have a high probability to share similar functions, cellular roles and sub-cellular locations. If two proteins have similar functions, they will theoretically share similar three-dimensional structures as well. Therefore, it is believed that if two proteins have similar secondary structures, they will also have similar three-dimensional structures and consequently share similar functions. As a result they will interact with each other. However, if these proteins have similar sequence, they do not always have similar secondary structures and consequently similar three-dimensional structures and functions. Based on these theories, we predict the interacting proteins in *Saccharomyces cerevisie* (baker's yeast) from the information of their secondary structures using computational method.

This paper proposes multiple independent fuzzy systems for predicting protein-protein interactions from the similarity of protein secondary structures. Our method consists of two main stages: (1) similarity score computation, and (2) similarity classification. The first stage involves three steps: (1) Multiple-sequence alignment (MSA)—finding multiple-sequence alignment for every family groups of proteins in *Saccharomyces cerevisie*, (2) Secondary structure prediction (SSP)—predicting secondary structure of aligned proteins sequence using secondary structure prediction tool called SSpro, and (3) Similarity measurement (Sim)—computing similarity scores of predicted secondary structures for every possible proteins pairs based on the number of three conformational states: helix (H), sheet (E), and coil (C).

In the classification stage, *N* multiple independent first order Sugeno Fuzzy Systems are generated to model the behavior of similarity scores of all possible proteins pairs to classify the interacting and non-interacting pairs; here *N* is the number of protein. Every system determines initial rules based on the clusters information obtained from the fuzzy clustering method. We employ principal component analysis in every system to compress the dimension of input data. Our model has been trained and tested using 1029 proteins with already known 2965 positive interactions of *Saccharomyces cerevisie* (baker's yeast). This proposed model achieves good accuracy when compared with experimentally determined proteins interactions from the Database of Interacting Proteins.

Keywords: *protein-protein interaction prediction, secondary structures, fuzzy system modeling.*

1. INTRODUCTION

Protein-protein interaction (PPI) is crucial for every organism. Most of the biological functions are mediated by protein interactions. Proteins may interact with each other for a long time to form protein complexes, a protein may be carrying another, or a protein may interact briefly with another protein just to modify it. Detecting which proteins interact, how they interact and what function is performed by their complex interaction is at least as important as predicting the three-dimensional structure of individual proteins. The information about such interactions improves our understanding of diseases and can provide the basis for new therapeutic approaches.

An impressive set of experimental approaches has been developed for the systematic analysis of protein interactions including yeast two-hybrid system, high-throughput, affinity chromatography, phage library display, and mass spectrometry methods (Tramontano, 2005; Mering et al, 2002; Ling et al, 2006). The yeast two-hybrid system works only with two-domain proteins in the yeast where the first domain's task is binding specific DNA sequences and the second domain is responsible for activating the transcription of a gene. In high-throughput technology, it allows the simultaneous analysis of thousands of parameters within a single experiment. For example, microarray analysis was developed to examine expression at the protein level to acquire quantitative and qualitative information about protein function.

During the 1990s, most of the methods focused on amino acids sequence comparison approaches for only completely sequenced genomes, such as *Helicobacter pylori*, *Bacillus subtilis*, *Mycoplasma genitalium* and others. Every gene of two different complete-sequenced bacteria, *H. Influenzae* and *E. Coli*, were clustered based on their functional classes in order to study the gene order relationships and genome organization in both bacteria (Tamames, 1997). The conservation of gene order method assumes that the proteins encoded by conserved gene pairs appear to interact physically. This method can also be used to predict functions of prokaryotic gene products (Dandekar *et al*, 1998). Another approach to predict PPIs is the gene fusion method that identifies gene-fusion events in complete genomes based on sequence comparison (Enright *et al*, 1999). The similarity of phylogenetic trees approach named as Mirrortree achieved 66% accuracy by considering the effects of the reference organisms and the identification of homologous proteins (Pazos *et al*, 2001; Sun *et al*, 2005). Besides that, a few more methods were proposed based on the similarity of phylogenetic trees, including partial correlation coefficient (Sato *et al*, 2003), intra-matrix correlations (Craig *et al*, 2007) and SVM-based method (Marangoni, 2003; Chen *et al*, 2005) with accuracies between 66 to 80%.

Different prediction approaches that exploit protein three-dimensional structures information have also been developed. For example, docking methods, threading-based methods and homology methods. Docking method has been developed by assuming that the putative interactors associate using the same interface patches as the seed interactors (Cockell *et al*, 2007). MULTIPROSPECTOR is a multimeric structure-based threading approach which aims to capture more distantly related or even analogous proteins (Lu *et al*, 2003). In homology methods, it is believed that protein-protein interaction can be modeled by known structures of protein complexes whose components are homologous or similar to other proteins whose interactions to be modeled (Szilagyi *et al*, 2005).

Although there are no concrete properties in predicting protein-protein interaction, it is experimentally verified that proteins with strong protein-protein interactions have a high probability of sharing similar functions, cellular roles, and/or sub-cellular locations. Therefore, if two proteins have similar functions, it is believed theoretically that they will also share similar three-dimensional structures. However, if two proteins have similar sequence, it is not strongly verified that they will also have similar function and interact with each other. Thus, it is believed that if two proteins have similar secondary structures, they will also have similar three-dimensional structures and therefore share similar functions and interact with each other.

1.1 Machine Learning Approaches for PPI Prediction

Machine learning approaches are best suited for problems where there is a large amount of data with unknown theoretical principles. In bioinformatics area, there are lots of problems that have lack of discovered theory, such as PPI prediction problem. Even though databases to give a variety of information for every protein are available, all the information cannot be fully exploited due to the lack of interaction theory yet.

Subsequent to the introduction of many machine learning approaches, Bock and Gough were among the pioneers that developed a method using Support Vector Machines (SVM) in PPI predicting. They proposed SVM-light to recognize and predict PPIs based on protein sequences and physico-chemical properties, i.e. charge and surface tension of protein (Bock *et al*, 2000). A kernel based on signature products method has

also been introduced to improve the accuracy in the range 70-80% by using 10-fold cross validation (Martin *et al*, 2005). Besides SVM, Hidden Markov models (HMMs) have been introduced to PPIs as well. HMMs were built with artificial multiple sequence alignment patches to search sequences with remote homology (Espadaler, 2005). However, the HMM-based methods do not achieve a good prediction compared to SVM mainly because of the lack of information on protein sequences used in HMMs.

In this work, a novel approach based on first order Sugeno fuzzy system is introduced to use secondary structure information of proteins to predict either stable or transient physical interactions among them. This paper is organized into five sections. The first section overviews PPIs, followed by the problem statement and proposed approach in the second and third sections, respectively. Section 4 provides a detailed discussion of the results. . The paper is concluded in section 5.

2. PROBLEM STATEMENT

The prediction of PPIs problem can be formulated as follows:

Given a set of amino acid sequences of any organism, $S = \{s_1, s_2, \ldots, s_N\}$ and a set of predicted secondary structure, $SS = \{ss_1, ss_2, \ldots, ss_N\}$ where N is the number of proteins, find the connected graph $G(V,E)$ where $V = \{p_1, p_2, \ldots, p_N\}$ represent a set of proteins and $E = \{w_{ij} \mid i, j = 1,2,\ldots,N\}$ is a set of similarity scores for connected proteins *i* and *j*.

Every predicted secondary structure can be presented in a sequence consists of secondary structure elements: helices (H), sheets (E) and coils (C). Every secondary structure element are presented as $ss_i = \{e_{i,1}, e_{i,2}, \ldots, e_{i,n}\}$ where *n* is a structure length.

The similarity score formula for proteins pair (i, j) can be written as below:

$$w_{ij} = \sum_{\alpha,\beta}^{n,m} \left(1 \quad if \quad e_{i,\alpha} = e_{j,\beta}\right) \tag{1}$$

with respect to $e_{i,\alpha} = e_{j,\beta}$ if elements match $H \to H$, $E \to E$, $C \to C$ or structure of coil match, $(H,E) \to C$ is satisfied. The global alignment procedure is applied here, where gaps will be added in the shorter fragment of $H \to H$, $E \to E$ or $C \to C$ matches. Note that, *n* and *m* are the lengths of secondary structure of proteins *i* and *j*, respectively.

3. METHOD

Our proposed model is a quantitative computational approach that consists of two main stages as shown in Figure1.

3.1 Similarity Score Computation

The first stage is to compute the similarity scores through the following steps.:

STEP 1: Multiple Sequence Alignment (MSA) - The first step of the method involves a multiple-sequence alignment to find the relationship among several sequences. All proteins in *S* were grouped according to their families by using a clustering protein sequences tool called CLUSS (Kelil et al, 2007). CLUSS clusters all the protein sequences based on matching amino acid subsequences. Proteins that belong to the same group/cluster must have higher sequence similarity compared with sequences from different groups. All sequences in every group were aligned by using go ur new multiple-sequence alignment method named the Rubber Band Technique (RBT) (Taheri et al, 2008). RBT is an

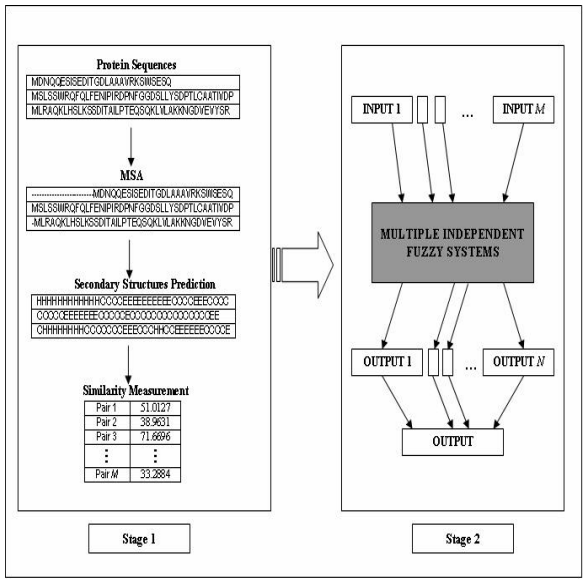

Figure 1. Framework of the proposed model for PPI prediction.

iterative heuristic technique used to solve the MSA problem. This technique is inspired by the natural behavior of a rubber band on a plate with several poles resembling location in input sequences that are most probably biologically related. This technique generated a Grid Answer Space (GAS) that is a multi-dimensional table to find relationship among the proteins to be aligned. The answer from RBT is a unique one arrowed line called the Rubber Band (RB) in this generated GAS. This RB generated the final alignment among proteins.

STEP 2: Secondary Structure Prediction (SSP) - The second step of the data set preparation involves secondary structure prediction. As mentioned earlier, databases of experimentally determined protein secondary structure are very limited (not all proteins have their secondary structure information in the databases). Therefore, SSpro (Cheng, 2005) as one of the most popular tools for secondary structure prediction is used in the proposed method. From the aligned sequences (results from MSA), SSpro predicts secondary structure for every protein. SSpro represents every element of secondary structure by three conformational states: H, E and C.

STEP 3: Similarity Measurement (Sim) - Based on these elements and a number corresponding to the length of the region, the method continues with the next step to measure the similarity of all pairs of proteins. For N proteins, we will have $N(N-1)$ possible interacting proteins. Similarity score for every pair is calculated from formula (1), where $w_{ij} = w_{ji}$ for proteins pairs (i, j) and (j, i). The scores are normalized to values in the range [0,100] where higher scores resemble higher similarity between two proteins.

3.2 Classification

In the second stage, we classify all the similarity scores using machine learning approach called first order Sugeno Fuzzy System. In this paper, we proposed a multiple independent fuzzy systems model to categorize the interacting proteins and non-interacting proteins from the given similarity scores of all possible neighbors for every protein.

Principal Component Analysis

Principal Component Analysis (PCA) is one of the tools in exploratory data analysis that involves mathematical procedure to transform large number of correlated variables into smaller uncorrelated variables. The uncorrelated variables are called principal components. Before calculating the principal components values, all the data must be standardized by using mean and standard deviation of every variable. The PCA transformation can be formulated as (2).

$$\begin{aligned} Y^T &= W^T X \\ &= V \sum X^T \end{aligned} \tag{2}$$

Where $V \sum X^T$ is the eigenvalue decomposition of covariance matrix of similarity scores matrix, W^T.

The first principal component considers as much of the variability in the data and the remaining variability is accounted for the other succeeding principal components as much as possible (Jolliffe, 2002). In other words, PCA is able to reduce the size of the input data and consequently reduce the complexity of the system. Therefore, we added PCA for every fuzzy system to compress all the N inputs into the M uncorrelated inputs where $M < N$, as shown in figure 2.

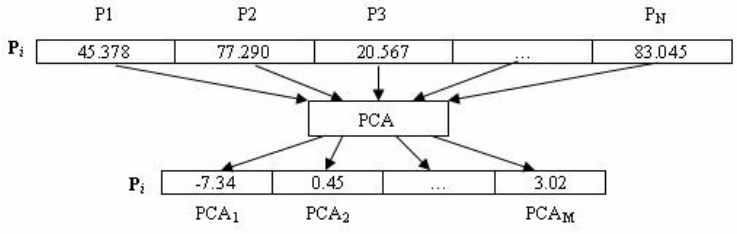

Figure 2. PCA transformation example.

Multiple Independent Fuzzy Model

Fuzzy system (FS) consisting of a set of fuzzy IF-THEN rules is used to map the system inputs to output. In fuzzy systems theory, the combination of different fuzzification and defuzzification functions with different rule base structures can lead to various solutions to a given task. However a single FS may not be suitable for large dimension dataset because it can possibly increase the complexity and consequently reduce the speed of the system (Yen et al, 1997; Cheng et al, 2002). Alternatively, the multiple fuzzy systems can be developed not only to speed up the whole systems but also to increase the reliability and simplicity of the system.

In this work, we construct a multiple independent FS with M inputs whose membership functions for every input are obtained from fuzzy clustering method (FCM). Inference rules for every subsystem are determined based on clusters from FCM. As a result, Gaussian membership functions with product inference rule were used at the fuzzification level. The associated membership function parameters were adjusted using a combination of backpropagation algorithm and least squares estimation during learning process. Our model has only one output in the range [0 1] for every system where higher scores resemble higher probability of interacting proteins.

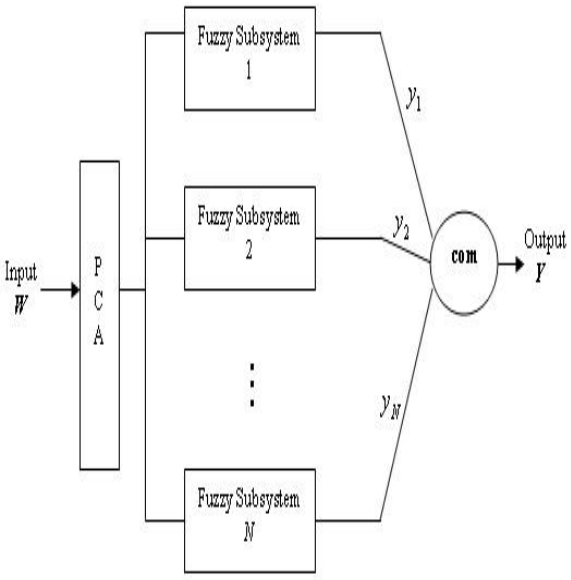

Figure 3. An architecture of multiple fuzzy model for protein-protein interaction prediction with an input matrix of similarity score, subsystem output, y_i for $i = 1,2,\ldots,N$ and "**com**" operation combines y_i by rows resulting in the final output Y.

After applying PCA to the input data, we have a smaller number of input data dimension, M as shown in Figure 2. All new input data are applied to every N independent fuzzy system. Every i-th fuzzy system classifies all possible links between protein i and all other proteins into interacting or non-interacting pairs by giving the output value in the range [0 1]. The collection of outputs from all N fuzzy systems will give an NxN matrix. Figure 3 shows the architecture of the proposed multiple independent fuzzy systems.

4. RESULTS AND DISCUSSION

The proposed model has been tested for 1029 *Saccharomyces cerevisie* (baker's yeast) proteins with known 2965 positive interactions among them. The positive interactions information was downloaded from the Database of Interacting Proteins (DIP) (Xenarios, 2000). DIP combines experimentally determined protein interactions information from various sources and it is updated on a regular basis.

During the first stage process, BLOSUM 62 scoring matrix and gap penalty of 5 and 1 for the gap opening and gap extension, respectively, were selected in RBT for MSA. We used random walk initialization mode for sequence length less than 200 and homogenous initialization mode, otherwise. RBT is executed ten times for every proteins group and its best result is considered as the final answer for MSA.

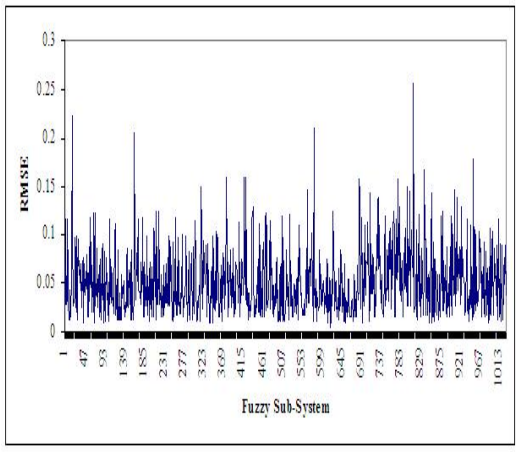

Figure 4. RMSE of 1029 subsystems.

In PCA process, we eliminate those principal components that contribute less than 1% to the total variation in the dataset. We used 10-fold cross validation test to evaluate the performance of our model. Every training and test data sets will be transformed separately. After the subsystem has been trained, the same transformation matrix would be used to transform the test dataset that are applied to the subsystem. PCA process has successfully transformed a 1029x1029 matrix dataset into a 1029x6 matrix. This situation

shows that among 1029 proteins, not all proteins have high connectivity with other proteins. Only 10% of these proteins have high connectivity with the maximum number of interaction is 77. After the validation test, our proposed model consists of $N = 1029$ subsystems and N different sets of inference rules (7 rules in average) with 0.0476419 of average of root mean square error (RMSE) for the whole model. Figure 4 shows the RMSE values for all fuzzy subsystems in our model. Only four subsystems achieve more than 0.2 of RMSE, while most of the remaining subsystems show good values of error with less than 0.05.

In this work, we also prepared seven different sizes of datasets which are 25, 50, 100, 200, 300, 400 and 1029 of proteins. We trained and tested our model with all the datasets to validate the stability and reliability of the model. As shown in figure 5, the positive success rate increases as the number of protein increases. However, there is a break down at dataset of 100 proteins. Our model predicts 73% of total known interacting proteins in dataset of 100 proteins. This happens because of the random selection of proteins that cause the fraction of negative interactions to be much larger than positive interactions. Other datasets achieve an average 80% of positive success rate with the highest rate at dataset of 1029 proteins. From 2965 known proteins interactions, our model correctly predicts 2290 protein interactions which is 85% positive success rate. The ROC curve of our proposed model shows a good figure of accuracy based on different cutoff values. The accuracy for the optimal cutoff of our model is 89% with 0.85 true positive rate (TPR) and 0.17 false positive rate (FPR). This pattern shows that our proposed model is stable and reliable for PPI prediction.

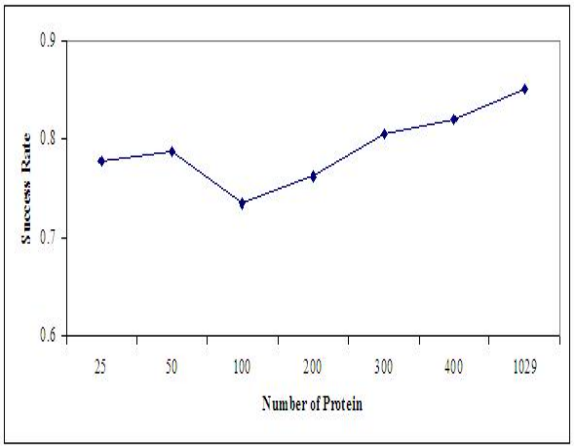

Figure 5. Positive success rates for different size of datasets.

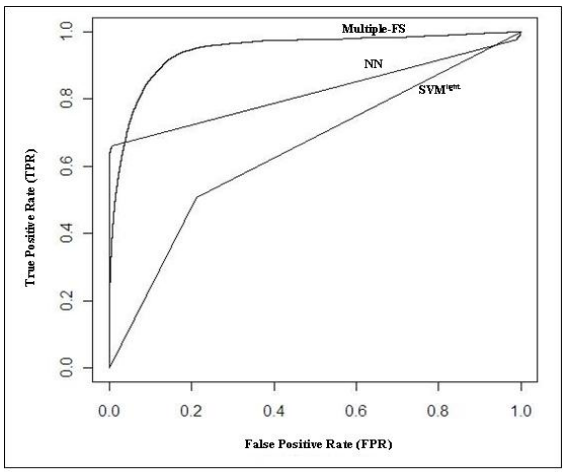

Figure 6. ROC curves of the proposed model, NN and SVM^{light}.

In addition, two machine learning methods were compared with the proposed method, which are SVM^{light} and Neural Network (NN) for dataset of 1029 proteins, as shown in Figure 6. SVM^{light} has been implemented by Bock (2002) while NN has never been applied for PPI prediction before. The same kernel function as in Bock (2002) was used in SVM^{light} to recognize the interacting pairs and non interacting pairs during 10-fold cross validation. Neural networks employed radial basis function with two layers and as many as N number of neurons. ROC curves of both methods show lower accuracies when compared to the multiple fuzzy systems and have similar pattern of the ROC curve.

In most experiments, the number of positive examples and negative examples are set to be in ratio 1:1. Unlike in our experiment, the consideration of all possible pairs of proteins makes our dataset much larger than other methods even with similar number of proteins. However, the proposed multiple fuzzy systems are able to specifically distinguish the positive and negative predictions with high sensitivity. When the same dataset was applied to other methods, such as NN and SVM^{light}, both methods couldn't achieve good accuracy as expected. Although they successfully predict the high number of true positive interactions, both methods predict high number of false positive interactions as well (similar pattern of ROC curve shown in Figure 6). This situation shows that SVM^{light} is limited to the small size dataset with the same number of positive and negative examples. Besides that, the fast training process to fit a smooth function (for NN) or to map training data to kernel space (for SVM^{light}) in both methods may cause the poor generalization of the classifier. Both methods took less than an hour to train the given data but our proposed method took three hours during training process for 1029 proteins dataset. However, the multiple fuzzy systems successfully generalized all the training data by achieving TPR 0.85 on validation data.

5. CONCLUSIONS

In this work, we proposed a model for protein-protein interaction prediction that employs the PCA process and multiple independent fuzzy systems. The proposed model predicts protein-protein interactions from the information of three conformational states of protein secondary structure. Our model achieved a good initial accuracy for 1029 proteins and we believe that it has better prediction accuracy for larger datasets. In the future, we will enhance our model with type-2 fuzzy system and increase more proteins information such as the co-localizations and functions annotations similarity.

REFERENCES

Bock, J.R. and Gough, D.A. (2000) Predicting protein-protein interactions from primary structure, *Bioinformatics*, **12**, 455-460.

Chen, X.W. and Liu, M. (2005) Prediction of protein-protein interactions using random decision forest framework, *Bioinformatics*, **21**, 4394-4400.

Cheng, C.B., C.J. Cheng, and E.S. Lee, *Neuro-Fuzzy and Genetic Algorithm in Multiple Response Optimization.* Computers and Mathematics with Applications, 2002. **44**: p. 1503-1514.

Cheng, J., Randall, A.Z., Sweredoski, M.J. and Baldi, P. (2005) SCRATCH: a protein structure and structural feature prediction server, *Nucleic Acids Research*, **33**, W72-W76.

Cockell, S.J., Oliva, B. and Jackson, R.M. (2007) Structure-based evaluation of in silico predictions of protein-protein interactions using comparative docking, *Bioinformatics*, **23**, 573-581.

Craig, R.A. and Liao, L. (2007) Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices, *BMC Bioinformatics*, **8**, 1-12.

Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact, *Trends Biochem Sci*, **23**, 324-328.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events, *Letters of Nature*, **402**, 86-90.

Espadaler, J., Romero-Isart, O., Jackson, R.M. and Oliva, B. (2005) Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships, *Bioinformatics*, **21**, 3360-3368.

Jolliffe, I.T., *Principal Component Analysis*. 2 ed. Springer Series in Statistics. 2002, New York: Springer.

Kelil, A., Wang, S., Brzezinsksi, R. and Fleury, A. (2007) CLUSS: Clustering of protein sequences based on a new similarity measure, *BMC Bioinformatics*, **8**.

Ling, C., Cho, Y.R., Hwang, W.C., Pei, P. and Zhang, A. (2006) Clustering methods in protein-protein interaction networks. In, *Knowledge Discovery in Bioinformatics*. John Wiley & Sons, Inc.

Lu, L., Arakaki, A.K., Lu, H. and Skolnick, J. (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale, *Genome Research*, **13**, 1146-1154.

Marangoni, F., Barberis, M. and Botta, M. (2003) Large scale prediction of protein interactions by a SVM-based method, *WIRN VIETRI 2003, LNCS*, 296-301.

Martin, S., Roe, D. and Faulon, J.-l. (2005) Predicting protein-protein interactions using signature products, *Bioinformatics*, **21**, 218-226.

Mering, C.V., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions, *Nature*, **417**, 399-403.

Pazos, F. and Valencia, A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction, *Protein Engineering*, **14**, 609-614.

Sato, T., Yamanishi, Y. and Horimoto, K. (2003) Prediction of protein-protein interactions from phylogenetic trees using partial correlation coefficient, *Genome Informatics*, **14**, 496-497.

Sun, J., Xu, J., Liu, Z., Liu, Q., Zhao, A. and Li, Y. (2005) Refined phylogenetic profiles method for predicting protein-protein interactions, *Bioinformatics*, **21**, 3409-3415.

Szilagyi, A., Grimm, V., Arakaki, A.K. and Skolnick, J. (2005) Prediction of physical protein-protein interactions, *Physical Biology*, **2**, S1-S16.

Taheri J, Zomaya AY, Zhou BB, "RBT-L: A Location Based Approach for Solving the Multiple Sequence Alignment Problem", The University of Sydney Technical Reports 2008, No. **626**.

Tamames, J., Casari, G., Ouzounis, C. and Valencia, A. (1997) Conserved clusters of functionally related genes in two bacterial genomes, *Journal of Molecular Evolution*, **44**, 66-73.

Tramontano, A. (2005) *The Ten Most Wanted Solutions in Protein Bioinformatics*. Chapman & Hall/CRC.

Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M. and Eisberg, D. (2000) DIP: The database of interacting proteins, *Nucleic Acids Research*, **28**, 289-291.

Yen, J., L. Wang, and R. Langari. *Multiple Fuzzy Systems for Function Approximation*. in *Annual Conference of the North American Fuzzy Information Processing Society - NAFIPS*. 1997: IEEE, Piscataway, NJ, United States.