JQMA 8(2) 2012, 9-20

FIS-PNN: A HYBRID COMPUTATIONAL METHOD FOR PROTEIN-PROTEIN INTERACTIONS PREDICTION USING THE SECONDARY STRUCTURE INFORMATION

(FIS-PNN: Suatu Kaedah Pengkomputeran Hibrid bagi Ramalan Interaksi Protein-Protein Menggunakan Maklumat Struktur Sekunder)

SAKHINAH ABU BAKAR[1], JAVID TAHERI[2] & ALBERT Y ZOMAYA[2]

ABSTRACT

The study of protein-protein interactions (PPI) is an active area of research in biology because it mediates most of the biological functions in any organism. This work is inspired by the fact that proteins with similar secondary structures mostly share very similar three-dimensional structures, and consequently, very similar functions. As a result, they must interact with each other. In this study we used our approach, namely FIS-PNN, to predict the interacting proteins in yeast from the information of their secondary structures using hybrid machine learning algorithms. Two main stages of our approach are similarity score computation, and classification. The first stage is further divided into three steps: (1) Multiple-sequence alignment, (2) Secondary structure prediction, and (3) Similarity measurement. In the classification stage, several independent first order Sugeno Fuzzy Inference Systems and probabilistic neural networks are generated to model the behavior of similarity scores of all possible proteins pairs. The final results show that the multiple classifiers have significantly improved the performance of the single classifier. Our method, namely FIS-PNN, successfully predicts PPI with 96% of accuracy, a level that is significantly greater than all other sequence-based prediction methods.

Keywords: protein-protein interaction prediction; hybrid method; secondary structure; machine learning algorithm

ABSTRAK

Interaksi protein-protein merupakan suatu bidang kajian biologi yang aktif kerana ia menjadi perantara bagi hampir semua fungsi biologi di dalam organisma. Kajian ini diilhamkan daripada hakikat bahawa protein yang mempunyai struktur sekunder yang serupa akan mempunyai struktur tiga-dimensi yang hampir serupa, dan seterusnya mempunyai fungsi yang sangat serupa. Oleh yang demikian, protein-protein tersebut akan berinteraksi di antara satu sama lain. Dalam kajian ini, digunakan pendekatan FIS-PNN untuk meramal interaksi di antara protein dalam ragi menggunakan maklumat struktur sekunder dan al-Khwarizmi hibrid pembelajaran mesin. Dua peringkat utama pendekatan ini adalah pengiraan skor keserupaan dan pengelasan. Peringkat pertama pula mempunyai tiga langkah: (1) Penjajaran multi-jujukan, (2) Peramalan struktur sekunder, dan (3) Pengukuran keserupaan. Dalam peringkat pengelasan pula, beberapa sistem pentaadbiran kabur Sugeno peringkat pertama yang tak bersandar dan rangkaian neural berkeberangkalian dijana untuk memodelkan telatah skor keserupaan bagi semua pasangan protein yang mungkin. Hasil kajian mendapati pengelas berbilang telah meningkatkan ketepatan ramalan berbanding dengan pengelas tunggal. FIS-PNN yang dicadangkan telah berjaya meramal interaksi protein-protein dengan ketepatan 96%, suatu tahap ketepatan yang jauh lebih baik berbanding dengan kesemua kaedah ramalan berasaskan jujukan yang lain.

Kata kunci: ramalan interaksi protein-protein; kaedah hibrid; struktur sekunder; al-Khwarizmi pembelajaran mesin

1. Introduction

Protein-protein interactions (PPIs) are crucial for every organism as most of the biological functions are mediated by them. In fact, 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 protein (Tramontano 2005). During the 1990s, because most of the PPI prediction methods were based on amino acids sequence comparisons, they were only applicable to complete sequenced genomes. For example, genes of two different complete-sequenced bacteria, *H. Influenzae* and *E. Coli*, were clustered based on their functional classes to investigate individual gene's relationship order (Tamames *et al.* 1997). Another approach to predict PPIs is 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 in the target organism (Pazos & Valencia 2001). Furthermore, a few more methods were proposed based on the similarity of phylogenetic trees, including partial correlation coefficient (Sato *et al.* 2003) and intra-matrix correlations (Craig & Liao 2007) with overall accuracies of 66-80%.

Subsequent to the introduction of many machine learning approaches, Bock and Gough were among the pioneers who managed to develop a method using Support Vector Machines (SVM) in PPI prediction. They proposed SVM-light to recognise and predict PPIs based on protein sequences and physico-chemical properties (Bock & Gough 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 also been introduced to PPIs. HMMs were built with artificial multiple sequence alignment patches to search sequences with remote homology (Espadaler *et al.* 2005).

Although there are no concrete properties in predicting PPI, it is experimentally verified that proteins with strong PPIs more probability share similar functions, cellular roles, and/or sub-cellular locations. Therefore, if two proteins have similar functions, it is theoretically believed that they also share similar three-dimensional structures (Tramontano 2005). These hypotheses can be used to conclude that if two proteins have similar secondary structures, they most probably have similar three-dimensional structures and therefore share similar functions and interact with each other.

2. Methods

In this paper, we introduced a new hybrid method that employed multiple independent fuzzy inference systems and probabilistic neural networks to predict PPI using the similarity score of proteins secondary structures. The proposed method, FIS-PNN, consists of two main stages as shown in Figure 1.

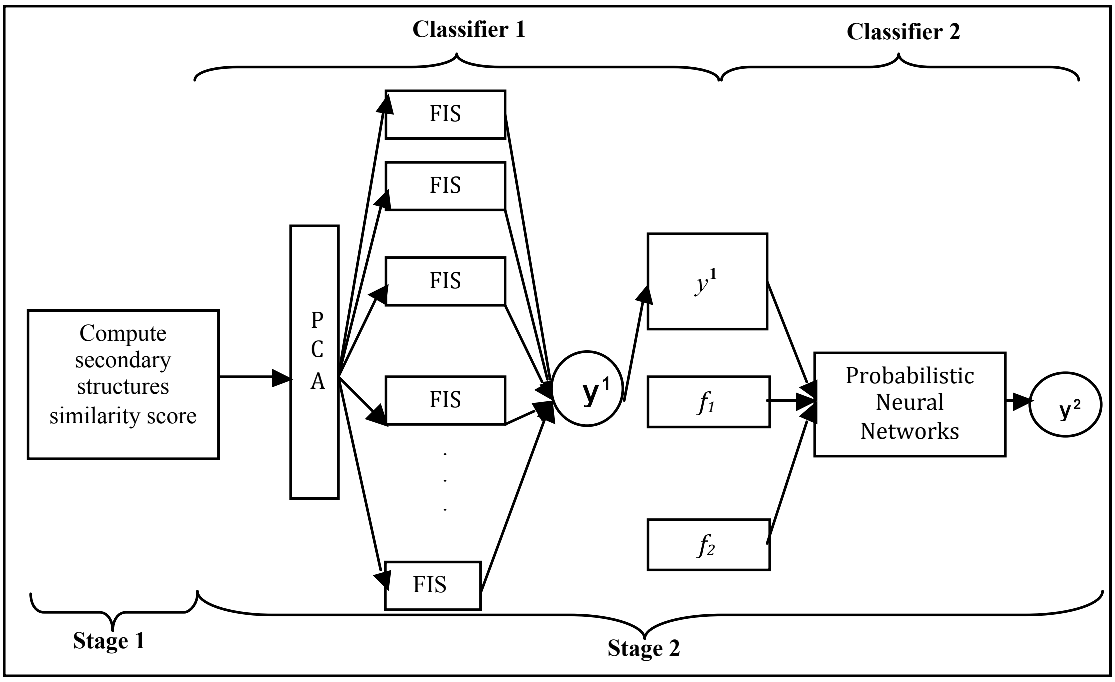

Classifier 1
Classifier 2
Compute secondary structures similarity score
P C A
FIS
FIS
FIS
FIS
FIS
y¹
y¹
f1
f2
Probabilistic Neural Networks
y²
Stage 1
Stage 2

Figure 1: Architecture of FIS-PNN

The prediction of PPI problem can be formulated as: given a set of amino acid sequences of any organism, $S = \{s_1, s_2, \ldots, s_N\}$ and their associated secondary structures, $SS = \{ss_1, ss_2, \ldots, ss_N\}$ 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 the structure length. In this case, the similarity score formula for proteins pair (i, j) can be written as $w_{ij} = \sum_{\alpha,\beta}^{n,m} (1\, if\ e_{i,\alpha} = e_{j,\beta})$ 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. Note that, *n* and *m* are the lengths of secondary structure of proteins *i* and *j*, respectively (Bakar *et al.* 2009).

2.1. *Secondary Structure Score*

The first stage is to compute the similarity scores through the following steps:
STEP 1: Multiple Sequence Alignment (MSA)
STEP 2: Secondary Structure Prediction (SSP)
STEP 3: Similarity Measurement (Sim)
More details regarding to this sub-section can be obtained from Bakar (2009).

2.2. *Hybrid Classification Method*

The FIS-PNN consists of two classifiers: (1) the multiple fuzzy inference systems (FIS), and (2) the probabilistic neural networks (PNN). The first classifier works with the similarity score for every possible protein pairs ($N \times N$ matrix) as the input. The output from the first classifier is an input of the second classifier. The final output is the output of the probabilistic neural networks that classify the given input into two classes.

2.2.1. *Fuzzy inference systems modeling*

Fuzzy inference systems (FISs) consist of a set of fuzzy IF-THEN rules to map a system's inputs to its outputs. In FISs theory, the combination of different fuzzification and defuzzification functions with different rule base structures can lead to various solutions to a given task (Taheri & Zomaya 2006). However, because a single FIS may not be suitable for large dimension datasets as it easily increases the complexity and reduces the speed of the system, multiple FISs are used instead. As a result, the whole system runs faster, becomes more reliable, and also much simpler.

In this work, we construct a set of independent FISs (FIS) with M inputs whose membership functions are obtained from fuzzy clustering method (FCM). Inference rules for every subsystem are determined based on clusters from FCM. Gaussian membership functions with product inference rule were used at the fuzzification level (Taheri & Zomaya 2006). The associated membership function parameters were set based on the combination of a backpropagation algorithm and a least squares estimation during the learning process. Our system has only one output in the range [0 1] for every system where higher scores resemble higher probability of interacting proteins.

After applying principal component analysis (PCA) to the input data, we have a smaller number of input data dimension, M. All new input data are applied to all N independent fuzzy systems where $M<N$. 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 is stored as an NxN matrix. Figure 2 shows the architecture of the proposed multiple independent fuzzy systems.

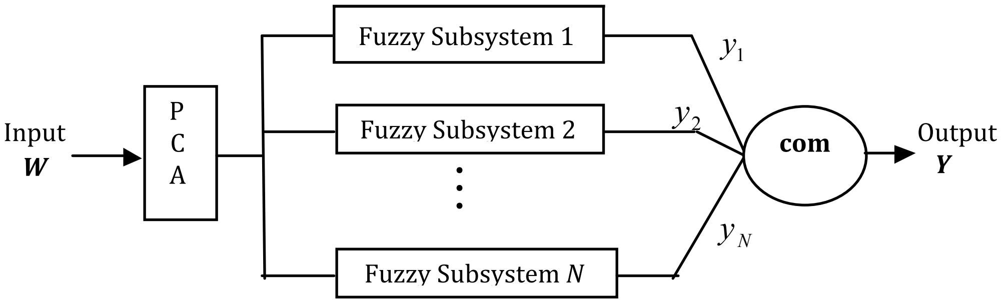


Figure 2: An architecture of multiple fuzzy inference systems model

2.2.2. *Probabilistic Neural Networks (PNN)*

An artificial neural network (ANN) is a computational model that simulates the structure of a neural network. The network consists of interconnected neurons and processes information. ANN is usually deployed to model complex or unknown relationships among inputs and outputs. The radial basis networks called probabilistic neural networks have been introduced by

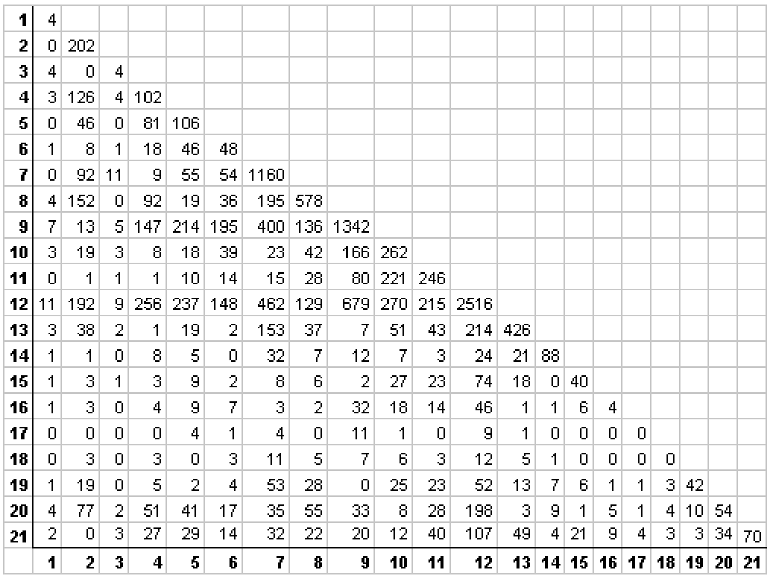
Specht (1990). The three layers PNN were developed to solve the classification problems where the radial basis function, e^{-x^2} is chosen as the transfer function for neurons in the hidden layer.

The PNN works as the second classifier in FIS-PNN. Output matrix of the first classifier, FIS is fed into the PNN as its input to generate the final symmetrical output matrix. Here, along receiving an input vector, the first layer computes distances from this input vector to all training input vectors; the second layer sums these contributions for each class of inputs to produce a vector of values between [0 1]. The output layer in this case classifies the vector by selecting the maximum value of the vector elements and produces a '1' for the chosen class and '0' for the other classes. In this study, the final output of the proposed method has two classes for interacting and non-interacting protein pairs.

2.3. *Computation of Protein Features*

Besides similarity scores of proteins' secondary structures, two other protein features (frequency of co-localisation, f_1, and similarity scores of function annotation, f_2) are also considered in predicting PPI. These features were added to FIS's output as a new input for second classifier, PNN, as shown in Figure 1.

A co-localisation matrix of known interacting proteins from Database of Interacting Protein (DIP) has been developed and compared with randomised interaction matrix (Xenarios I *et al.* 2000). The resulting matrix called the co-localisation weight matrix, ***L*** divides sub-cellular localisation into 21 categories as shown in Figure 3. It is known that most proteins move through several sub-cellular localisations in ***L***. Therefore, we compute the frequency of co-localisation for every protein pairs using the following formula $f_1\left(p_1,p_2\right)=\underset{\forall L_i\in p_1,\forall L_j\in p_2}{MAX}\left\{Localization\left(L_i,L_j\right)\right\}$ (Lee *et al.* 2005), where $Localization(\cdot)$ refers to combination of localisation weight in ***L***.



1	4																				
2	0	202																			
3	4	0	4																		
4	3	126	4	102																	
5	0	46	0	81	106																
6	1	8	1	18	46	48															
7	0	92	11	9	55	54	1160														
8	4	152	0	92	19	36	195	578													
9	7	13	5	147	214	195	400	136	1342												
10	3	19	3	8	18	39	23	42	166	262											
11	0	1	1	1	10	14	15	28	80	221	246										
12	11	192	9	256	237	148	462	129	679	270	215	2516									
13	3	38	2	1	19	2	153	37	7	51	43	214	426								
14	1	1	0	8	5	0	32	7	12	7	3	24	21	88							
15	1	3	1	3	9	2	8	6	2	27	23	74	18	0	40						
16	1	3	0	4	9	7	3	2	32	18	14	46	1	1	6	4					
17	0	0	0	0	4	1	4	0	11	1	0	9	1	0	0	0	0				
18	0	3	0	3	0	3	11	5	7	6	3	12	5	1	0	0	0	0			
19	1	19	0	5	2	4	53	28	0	25	23	52	13	7	6	1	1	3	42		
20	4	77	2	51	41	17	35	55	33	8	28	198	3	9	1	5	1	4	10	54	
21	2	0	3	27	29	14	32	22	20	12	40	107	49	4	21	9	4	3	3	34	70
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**

Figure 3: The co-localisation weight matrix with 21 categories of sub-cellular localisations

The similarities of functional annotation are computed based on a hierarchical tree structure called FunCat (Ruepp *et al.* 2005). FunCat has 28 main functional categories with up to six levels of increasing specificity. The similarity score of functional annotation for every proteins pair are computed as $f_2\left(p_1, p_2\right) = \underset{\forall f_i \in p_1, \forall f_j \in p_2}{MAX} \left\{2^{LCA\left(f_i, f_j\right)}\right\}$ (Lee *et al.* 2005), where LCA is the lowest common ancestor of the two proteins.

3. Results and Discussion

The FIS-PNN has been tested using 1029 yeast proteins with 2965 already known positive interactions among them. The positive interactions information was downloaded from the DIP (Xenarios *et al.* 2000). During the first stage of FIS-PNN, BLOSUM62 scoring matrix with gap opening penalty of 5 and gap extension penalty of 1 were selected in RBT for MSA (Taheri & Zomaya 2010). We used random walk initialisation mode for sequence length less than 200 and homogenous initialisation mode, otherwise. RBT is executed ten times for every proteins group and its best result is considered as the final answer for MSA.

For the second stage, we executed our first classifier, FIS with similarity scores obtained from the first stage as inputs. PCA eliminate those principal components that contribute less than 1% to the total variation in the input vector. We used 10-fold cross validation test to evaluate the performance of our classifiers. After a subsystem is trained, the same transformation matrix is used to transform the test dataset that are applied to the subsystem. PCA process has successfully transformed a 1029×1029 matrix dataset into a 1029×6 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 77 interactions. After the validation test, our first classifier consists of N = 1029 subsystems and N different sets of inference rules (7 rules in average) with 0.0476 of average of root mean square error (RMSE).

Then we used results from FIS as an input for the probabilistic neural networks. Since the results matrix is symmetrical, we decided to use only the lower triangular as input for the second classifier. In this study, we executed PNN with four different inputs as C1, C2, C3, and C4. C1 consisted of FIS's output, C2 consisted of FIS's output plus f_1, C3 consisted of FIS's output plus f_2, and C4 consisted of both protein features, i.e., FIS's output, f_1, and f_2. PNN has been trained by gradient descent with adaptive learning rate back-propagation and tested using 10-fold cross validation.

Due to different protein information used in FIS-PNN and other published PPI prediction methods, it is impossible to do performance comparisons with all these methods. These methods are only applicable with their own datasets. For example, Mirrortree approach requires information of protein sequence elements for yeast and other reference organisms. Same goes to other methods, where major changes are needed before applying our datasets. However, in this study we picked two of the best known algorithms, SVM-light from Bock (2000) and *k*-nearest neighbour method, to gauge the performance of our novel approach, FIS-PNN.

3.1. *Single Classifier versus Multiple Classifiers*

Our proposed hybrid classification method, FIS-PNN, that is consists of two different classifiers was able to successfully classify interacting and non-interacting proteins based on only their secondary structure similarity. The first classifier achieve 85% of true positive rate while the second classifier improved the classification performance by correctly predict 2791 interacting proteins; that is, 91% of accuracy without addition any extra protein's feature.

In our experiment, two machine learning methods were compared with FIS-PNN: SVM-light and *k*-nearest neighbor (KNN), as well as their combination. SVM-light has been implemented by Bock (2000) and *k*-nearest neighbour has never been applied for PPI prediction before. The same kernel function as in Bock (2000) was used in SVM-light to recognise the interacting pairs and non interacting pairs during 10-fold cross validation. *K*-nearest neighbour employed Euclidean distance as a distance metric with $k = 1$.

As proven in Thomas (2000), our study demonstrates that combination of multiple classifiers can always perform better than individual classifiers. Figure 4 shows that the probability neural networks achieves the best performance among other single machine learning classifier methods with 74.57% of accuracy, while SVM-light achieves lower accuracy and KNN was failed to classify the given input data. Although SVM-light successfully predicts the high number of true positive interactions, it predicts high number of false positive as well. This situation shows that SVM-light is limited to the small-sized datasets with the fairly equivalent number of positive and negative links. However, both methods (SVM-light and KNN) significantly improve their performance when combined with FIS to achieve 83.9% and 84.05% of accuracy, respectively (Figure 5). It is shown that the multiple classifiers method has a capability to improve the performance of single classifier method. In this study, our proposed FIS is able to make significant classification improvements as it can be observed through the wide area between multiple classifiers and single classifier ROCR curves in Figure 5.

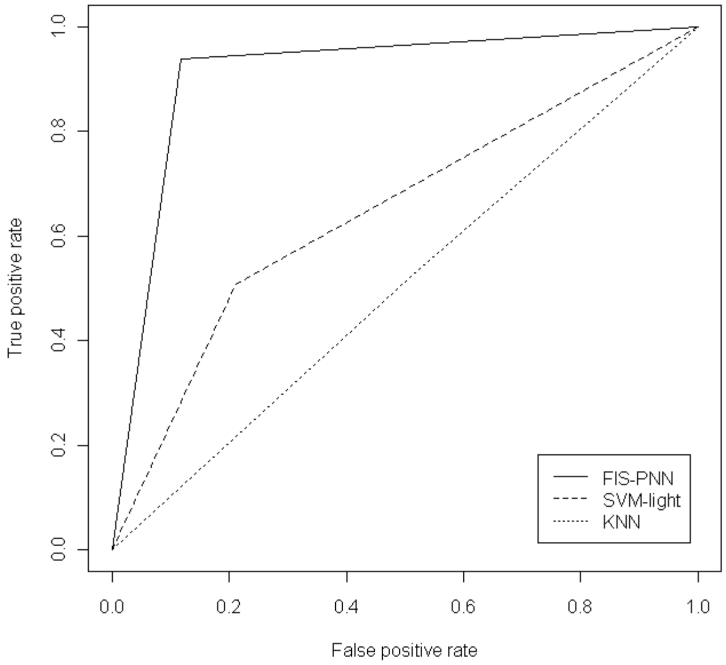
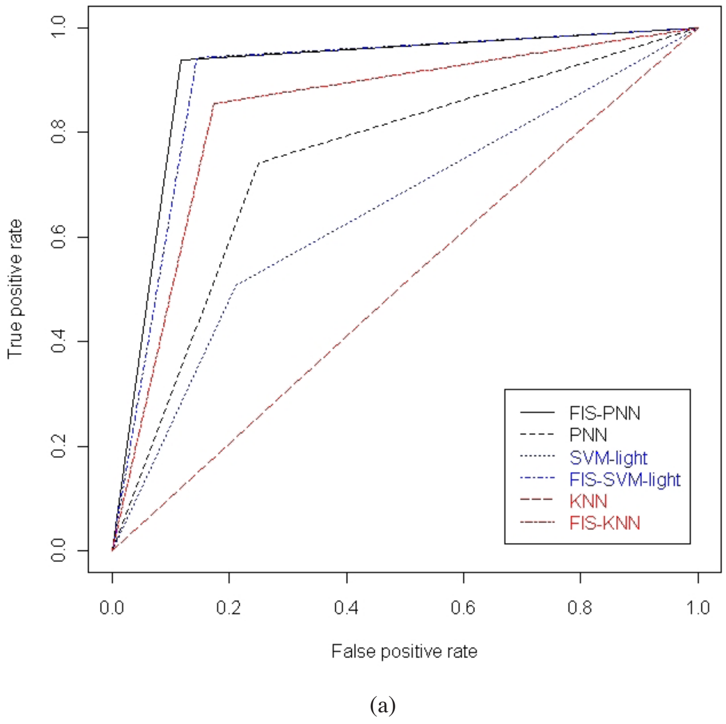


Figure 4: Performance comparison of FIS-PNN, SVM-light, and KNN



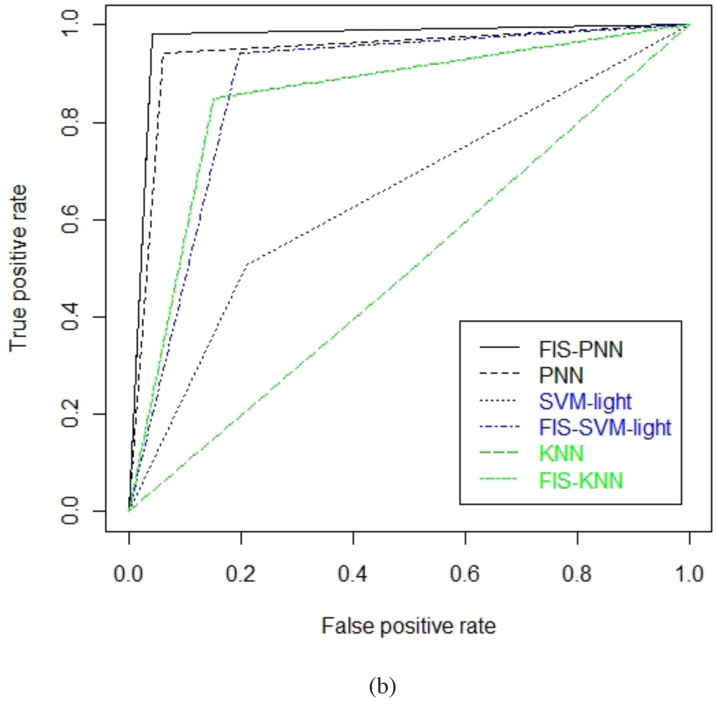
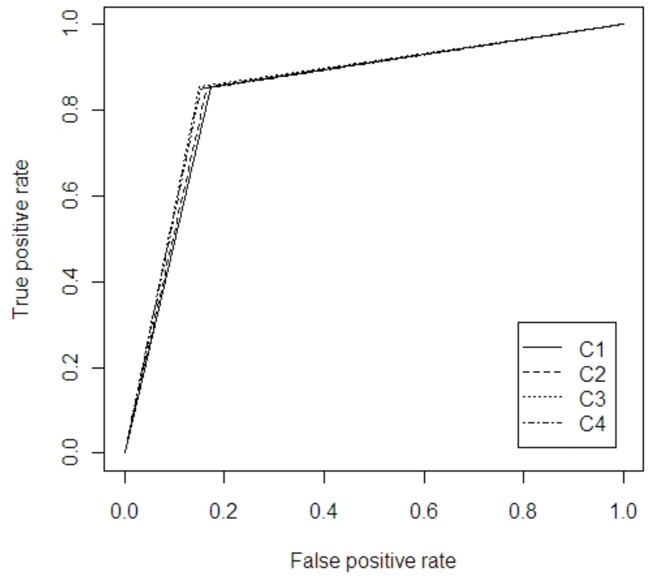


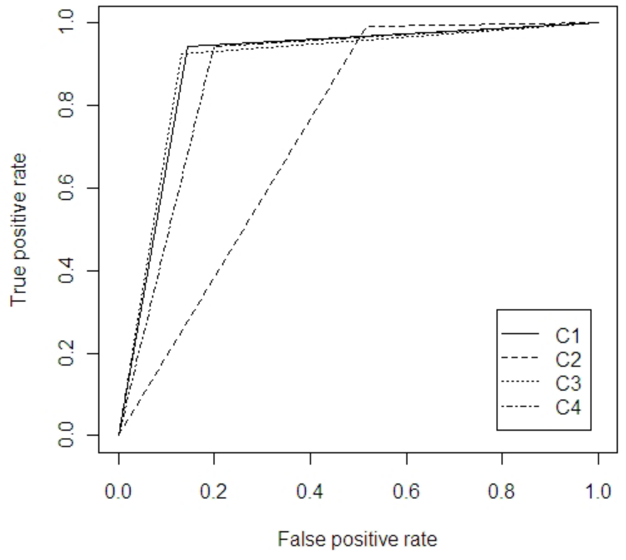
Figure 5: The ROC curves for all methods using (a) similarity scores of proteins' secondary structures, and (b) similarity scores of proteins' secondary structures plus other proteins' features

3.2. *Protein Features*

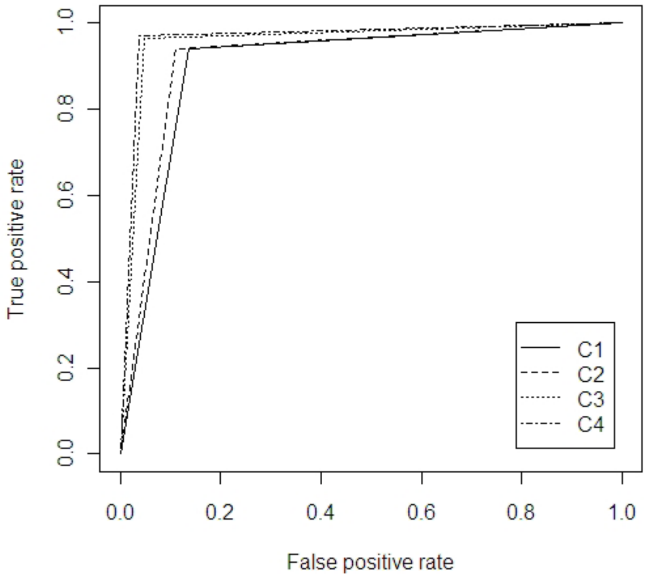
In this study, besides using the similarity score of proteins secondary structures, we also considered two other protein features (similarity of functional annotation and frequency of co-localisation) for every protein pair. We developed four different datasets to be trained and tested using FIS-PNN. Figure 6 shows results of these experiments with different input sets for the second classifier, (PNN, SVM-light and KNN).



(a)



(b)



(c)
Figure 6: FIS-PNN's performance for different datasets

Although a protein's secondary structure reveals invaluable information regarding the protein's function, it might not be adequate for an accurate prediction. Therefore, other proteins' features are considered in this work to significantly improve accuracy of PPI predictions. This yields to 97% of accuracy when all proteins' features are considered. Figure 6(c) shows reducing of false positive rate when more protein features are considered. The figure also shows that information from proteins' secondary structures and their functional annotations are much more coherent in predicting PPIs when compared to information of protein co-localisations. Similar results are observed when FIS-KNN is applied to the same datasets as shown in Figure 6(a). FIS-SVM-light (Figure 6(c)), on the other hand, shows opposite results as its performance decreases for larger datasets. This could be a result of the fast training process in this technique to map training data to kernel space.

3.3. *Balanced versus Imbalanced Dataset*

In most experiments, the numbers of positive and negative interactions are set to have a 1:1 ratio. The different sizes of datasets and the ratio of their positive and negative interactions are in fact the main factors that affect the performance of a classification method. In this work, we have also tested the performance of FIS-PNN with both balanced and imbalanced datasets. Here, imbalanced dataset had the ratio of 1:2, i.e., the number of negative interactions was twice the number of positive ones.

Table 1 shows results of our tests and reveals that accuracy of our propose algorithm, FIS-PNN, is not heavily affected by the level of asymmetry in the datasets. FIS-PNN was also capable of successfully achieving high sensitivity and specificity for various sizes of datasets without great dependency on their asymmetry ratio. FIS-KNN also successfully differentiates interacting proteins and non-interacting proteins in imbalanced dataset with accuracy between

83-85%. Unlike FIS-KNN, FIS-SVM-light could not perform accurate classification for imbalanced dataset. This method achieves low specificity for imbalanced datasets C1 and C2, high specificity for imbalanced datasets C3 and C4.

Table 1: The general performance of PPI prediction methods

	Balance				**Imbalance**			
	C1	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
FIS-PNN								
Accuracy	0.911	0.93	0.969	0.97	0.8951	0.9102	0.9545	0.9645
Sensitivity	0.939	0.942	0.975	0.982	0.8482	0.8654	0.9312	0.941
Specificity	0.883	0.919	0.963	0.958	0.9185	0.9325	0.9661	0.9762
FIS-KNN								
Accuracy	0.8405	0.8428	0.8516	0.8491	0.8337	0.8381	0.8535	0.8527
Sensitivity	0.8546	0.8509	0.8543	0.8496	0.7774	0.7781	0.7841	0.7777
Specificity	0.8263	0.8347	0.8489	0.8486	0.8619	0.8681	0.8882	0.8902
FIS-SVM-light								
Accuracy	0.839	0.7775	0.8899	0.8723	0.5892	0.7432	0.8742	0.8693
Sensitivity	0.9602	0.9815	0.9297	0.9396	0.9907	0.9868	0.9218	0.9366
Specificity	0.7177	0.5735	0.8501	0.8049	0.3884	0.6214	0.8504	0.8356

4. Conclusions

In this study, we deployed similarity scores of proteins' secondary structures as a new domain of protein information to predict accurate PPI. We predict PPIs based on formations helices, coils and sheets in proteins' secondary structure using our proposed classification method, FIS-PNN. The FIS-PNN successfully predicts PPIs with 96% of accuracy to significantly outperform many existing sequence-based prediction methods. This proposed method is able to efficiently classify large datasets using either solely information of their secondary structures, or in addition to other protein features. Results articulate feasibility of our approach even in formidable cases of imbalanced datasets with a large number of negative interactions.

References

Bakar S. A., Taheri J., & Zomaya A. Y. 2009. Fuzzy Systems Modeling for Protein-protein Interaction Prediction in Saccharomyces Cerevisie. *18th World IMACS / MODSIM Congress: Australia*, pp. 782-788.

Bock J. R. & Gough D. A. 2000. Predicting Protein-protein Interactions from Primary Structure. *Bioinformatics* 12(5): 455-460.

Craig R. A. & Liao L. 2007. Phylogenetic Tree Information Aids Supervised Learning for Predicting Protein-protein Interaction based on Distance Matrices. *BMC Bioinformatics* **8**(6): 1-12.

Enright A. J., Iliopoulos I., Kyrpides N. C. & 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. & Oliva B. 2005. Prediction of Protein-protein Interactions using Distant Conservation of Sequence Patterns and Structure Relationships. *Bioinformatics* **21**(16): 3360-3368.

Lee M. S., Park S. S. & Kim M. K. 2005. A Protein Interaction Verification System Based on a Neural Network Algorithm. *IEEE Computational Systems Bioinformatics Conference Workshops*. IEEE Computer Society; 151-154.

Martin S., Roe D. & Faulon J. 2005. Predicting Protein-protein Interactions using Signature Products. *Bioinformatics* **21**(2): 218-226.

Pazos F. & Valencia A. 2001. Similarity of Phylogenetic Trees as Indicator of Protein-protein Interaction. *Protein Engineering* **14**(9): 609-614.

Ruepp A., Zollner A., Maier D., Albermann K., Hani J., Mokrejs M., Tetko I., Guldener U., Mannhaupt G. &

Munsterkotter M. 2004. The FunCat: A Funtional Annotation Scheme for Systematic Classification of Proteins from Whole Genomes. *Nucleic Acids Research* **32**(18): 5539-5545.

Sato T., Yamanishi Y. & Horimoto K. 2003. Prediction of Protein-protein Interactions from Phylogenetic Trees using Partial Correlation Coefficient. *Genome Imformatics* **14**: 496-497.

Specht D. F. 1990. Probabilistic Neural networks and the Polynomial Adaline as Complementary Techniques for Classification. *IEEE Transactions on Neural Networks* **1**(1): 111-121.

Taheri J. & Zomaya A. Y. 2006. Fuzzy Logic. In *Handbook of Nature-Inspired and Innovative Computing.* Edited by Zomaya A. Y.: New York: Springer Science + Business Media Inc.

Taheri J. & Zomaya A. Y. 2010. RBT-L: A Location Based Approach for Solving the Multiple Sequence Alignment Problem. *International Journal of Bioinformatics Research and Applications (IJBRA)* **6**: 37-57.

Tamames J., Casari G., Ouzounis C. & Valencia A. 1997. Conserved Clusters of Functionally Related Genes in Two Bacterial Genomes. *Journal of Molecular Evolution* **44**: 66-73.

Thomas G. D. 2000. Ensemble Methods in Machine Learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems.* Springer-Verlag.

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. & Eisberg D. 2000. DIP: The Database of Interacting Proteins. *Nucleic Acids Research* **28**(1): 289-291.

[1]Pusat Pengajian Sains Matematik
Fakulti Sains dan Teknologi
Universiti Kebangsaan Malaysia
43600 UKM Bangi
Selangor DE, MALAYSIA
*E-mail: sakhinah@ukm.my**

[2]School of Information Technologies
Faculty of Engineering and IT
The University of Sydney
NSW 2006, AUSTRALIA.
E-mail: javid.taheri@sydney.edu.au, albert.zomaya@sydney.edu.au

* *Corresponding author*