Proceedings of the 5th WSEAS Int. Conf. on APPLIED INFORMATICS and COMMUNICATIONS, Malta, September 15-17, 2005 (pp313-318)

Improving Prediction Accuracy for Protein Structure Classification by Neural Network Using Feature Combination

Ken-Li Lin[a,b], Chun Yuan Lin[c], Chuen-Der Huang[d], Hsiu-Ming Chang[e], Chiao Yun Yang[f], Chin-Teng Lin[a], Chuan Yi Tang[f], D.Frank Hsu[g,*]

[a]Department of Electrical and Control Engineering,
National Chiao Tung University, Hsinchu, Taiwan
[b]Computer Center of Chung Hua University, Hsinchu,Taiwan
[c]Institute of Molecular and Cellular Biology, [e]Brain Research Center, [f]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
[d]Department of Electrical Engineering, Hsiuping Institute of Technology, Taichung, Taiwan
[g]Department of Computer and Information Science,
Fordham University, New York, NY 10023, USA

Abstract: - The classification of protein structures is essential for their function determination in bioinformatics. At present time, a reasonably high rate of prediction accuracy has been achieved in classifying proteins into four classes in the SCOP. However, it is still a challenge for classifying proteins into fine-grained folding categories, especially when the number of possible folding patterns as those defined in the SCOP is large. In our previous work, we have proposed a hierarchical learning architecture (HLA), two indirect coding features, and a gate function to differentiate proteins according to their classes and folding patterns. Our prediction accuracy rate for 27 folding categories was 65.5% compared favorably to previous results by Ding and Dubchak with 56.5% prediction accuracy rate. The success of the protein structure classification depends on two factors: the computational methods used and the features selected. In this paper, we use a combinatorial fusion analysis technique to facilitate feature selection and combination for improving predictive accuracy in protein structure classification. When applying the combinatorial fusion to our previous work, the resulting classification has an overall prediction accuracy rate of 87.8% for four classes and 70.9% for 27 folding categories. These rates are significantly higher than our previous work and demonstrate that combinatorial fusion is a valuable method for protein structure classification.

Key-Words: - machine learning; neural network; rank function; score function; rank/score functions; diversity graph.

1 Introduction

Large-scale sequencing projects produce a massive number of proteins with putative amino acid sequences but much less is known in terms of their three dimensional structure. Some famous structure databases, such as the Structural Classification of Proteins (SCOP) [4], contribute only no more than 32000 entries in the Protein Data Bank (PDB) (PDB: 31971 entries in 26-Jul. 2005). This is only about 20% of collections in the Swiss-Port (Swiss-Port release version 47.5: 188477 entries in 19-Jul. 2005). Therefore, to extract structural information just from the sequence databases becomes an important issue.

Previous research [1] have shown that an accuracy rate of 70-80% has been achieved to classify most of proteins into four classes according to their amino acid sequence information (i.e., all-alpha (all-α), all-beta (all-β), alpha/beta (α/β), and alpha+beta ($\alpha+\beta$)). However, less optimal results are obtained if a more complicated category is used such as the one with protein folding patterns [5].

Ding and Dubchak [5] proposed a taxonmetric approach for protein folding classification (into 27 folding patterns) beyond four simple classes with a Neural Network (NN) and Support Vector Machine (SVM) [15]. They predicted protein folds according to six single-parameter features 'C', 'S', 'H', 'P', 'V', and 'Z' first, then a combinatorial multiple-parameter features were formed and used in protein folding classification. They demonstrated that one feature 'CSHP' had the highest overall prediction accuracy rate for 27 folding categories at 56.5% by SVM.

In our previous work [10], extra features and a gate function were defined. We proposed two additional indirect coding features 'B' and 'SB' to correlate 'neighboring' di-peptide pairs with protein

structure classification. Then, two new features 'CSHPVZ+B' and 'CSHPVZ+B+SB' are formed to classify protein folding patterns. Due to the large number of input dimensions for these features, we used a gate function 'G' to reduce input dimensions of them first and then formed three new features 'CSHPVZ+G', 'CSHPVZ+B+G', and 'CSHPVZ+B+SB+G'. In addition to NN and SVM, we also constructed a new computational architecture called hierarchical learning architecture (HLA). In HLA, a protein is classified into one of four classes at first, and then further classified into one of 27 folding patterns. With the feature 'CSHPVZ+B+SB', we improved the prediction accuracy rate for 27 folding categories by 9%, compared with the result from Ding and Dubchak [5].

In this paper, we apply the technique of combinatorial fusion [6, 8, 9, 18] not only for better protein structure classification, but also for better feature selection and combination. In combinatorial fusion, results from various features are combined to obtain predictions with higher accuracy rate. We start with eleven features to assign protein class and folding patterns. Then, some explicit rules from combinatorial fusion in information retrieval (IR) and virtual screening (VS) [6, 8, 9, 13, 18] are used together with a special diversity rank/score graph to choose the best discriminating features for further combination. The proposed rules for proper feature selection are to reduce the complexity at the beginning. Then, we systematically choose the best discriminating features according to the diversity of these features, which is represented in a diversity rank/score graph. Our experimental results achieves an overall prediction accuracy rate at 87.8% for predicting protein classes and 70.9% for predicting protein folding patterns which are higher than our previous work at 83.6% and 65.5%, respectively.

2 Computational Framework and Architecture

2.1 Protein Data Sets

We use the data sets from Ding and Dubchak [5] which were originated from the SCOP for training and testing. Any pair of two proteins in the training data set is less than 35% identical in any aligned subsequence longer than 80 residues. All proteins in the testing data set are less than 40% identical to each other. No protein in the testing data set is more than 35% identical to any protein in the training data set. The number of proteins for training and testing data set is 313 and 385, respectively. Table 1 shows the number of proteins in different classes and folding patterns used in this paper.

Table 1. The variety in protein structures for training and testing

Classes	Folding patterns	No. of proteins (Training)	No. of proteins (Testing)
1. all-α	1. α_1: Globin-like	13	6
	2. α_2: Cytochrome *c*	7	9
	3. α_3: DNA-binding 3-helical bundle	12	20
	4. α_4: 4-helical up-and-down bundle	7	8
	5. α_5: 4-helical cytokines	9	9
	6. α_6: Alpha; EF-hand	7	9
2. all-β	7. β_1: Immunoglobulin-like β-sandwich	30	44
	8. β_2: Cupredoxins	9	12
	9. β_3: Viral coat and capsid proteins	16	13
	10. β_4: ConA-like lections/glucanases	7	6
	11. β_5: SH3-like barrel	8	8
	12. β_6: OB-fold	13	19
	13. β_7: Trefoil	8	4
	14. β_8: Trypsin-like serine proteases	9	4
	15. β_9: Lipocalins	9	7
3. α/β	16. $(\alpha/\beta)_1$: (TIM)-barrel	29	48
	17. $(\alpha/\beta)_2$: FAD (also NAD)-binding motif	11	12
	18. $(\alpha/\beta)_3$: Flavodoxin-like	11	13
	19. $(\alpha/\beta)_4$: NAD(P)-binding Rossmann-fold	13	27
	20. $(\alpha/\beta)_5$: P-loop containing nucleotide	10	12
	21. $(\alpha/\beta)_6$: Thioredoxin-like	9	8
	22. $(\alpha/\beta)_7$: Ribonuclease H-like motif	10	14
	23. $(\alpha/\beta)_8$: Hydrolases	11	7
	24. $(\alpha/\beta)_9$: Periplasmic binding protein-like	11	4
4. $\alpha+\beta$	25. $(\alpha+\beta)_1$: β-grasp	7	8
	26. $(\alpha+\beta)_2$: Ferredozin-like	13	27
	27. $(\alpha+\beta)_3$: Small inhibitors, toxins, lectins	12	27

2.2 Features

Different features may result in different classifications. Ding and Dubchak [5] proposed six single-parameter features based on physical, chemical, and structural properties of the constituent amino acids for protein structure classification. These features are amino acid composition (C), predicted secondary structure (S), hydrophobicity (H), normalized van der Waals volume (V), polarity (P), and polarizability (Z). Five multiple-parameter features, 'CS', 'CSH', 'CSHP', 'CSHPV', and 'CSHPVZ' were also constructed to classify protein folding patterns. They finally determined one feature 'CSHP' with the highest overall accuracy rate for protein structure prediction with SVM.

In our previous work [10], we used the N-gram concept to propose two indirect coding features, generated from the bigram (B) and the spaced-bigram coding (SB) scheme. We combined the six features proposed by Ding and Dubchak [5] and our two features to form two new features 'CSHPVZ+B' and 'CSHPVZ+B+SB'. Due to the large number of input dimensions for these features, we used a gate function 'G' to reduce the input dimensions of them and then formed three new features 'CSHPVZ+G', 'CSHPVZ+B+G', and 'CSHPVZ+B+SB+G'. By comparing these features above, the time of training

and testing for a certain feature with 'G' was much less than that for that without 'G'. However, each of these features with 'G' lost about 3% prediction accuracy rate by comparing to it without 'G'. We showed that using the feature 'CSHPVZ+B+SB' together with NN outperformed all features used by Ding and Dubchak [5] in terms of prediction accuracy rate for protein structure classification.

2.3 Computational Architecture

The NNs have been commonly used in many fields, such as input-output mapping and bioinformatics [16]. We use NN as a multi-class classifier to build HLA. The Radial Basis Function Network (RBFN) is a three-layer network with Gaussian function that is suitable to be a classifier [8]. Hence, we adopted the RBFN model in this paper. The HLA [10] consists of a two-level procedure. In the first level, a protein is classified into one of four classes by a multi-class classifier (classifier 1 in Fig. 1). Then, in the second level, it is further classified into one of f_i folding patterns by the corresponding multi-class classifier (f_1, f_2, f_3, and f_4 is equal to 6, 9, 9 and 3 in classifier 2, 3, 4, and 5 respectively in Fig. 1). In the current work, we incorporated combinatorial fusion in HLA for the testing data set, as shown in Fig. 1. For the training data set, HLA is used without combinatorial fusion. To predict which of four classes a protein belongs to with HLA, we use eleven features to assign class to each protein in the testing data set at first. Then, we use the technique of combinatorial fusion to select the best feature and to combine results for the protein class discrimination. Finally, the protein class is predicted with the combined feature. For protein folding patterns within each protein class, combinatorial fusion is applied again for feature selection and combination in order to predict protein folding patterns.

3 Combinatorial Fusion and Diversity Graph

Our approach to combination methods and feature selection in protein structure classification is analogous to those used in IR [8, 9, 13], pattern recognition [17], molecular similarity searching and VR [18], and microarray gene expression analysis [2, 3, 11]. Moreover, we adopt some of the notations and terminologies from [9] and [6].

When a protein sequence is given and a feature A is considered, let $s_A(x)$ be a function that assign a real number to the class (or folding pattern) x in the set of all n classes (or folding patterns) $D = \{c_1, c_2, ..., c_n\}$. We view the function $s_A(x)$ as the score function with

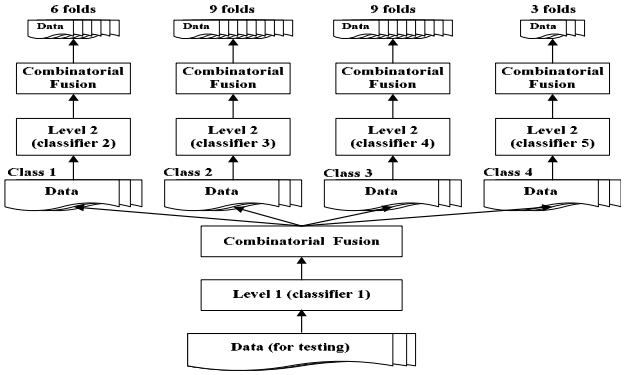


Fig. 1. The architecture of HLA together with combinatorial fusion

respect to the feature A from D to R (the set of real numbers). When treating $s_A(x)$ as an array of real numbers, it would become a rank function $r_A(x)$ after sorting the $s_A(x)$ array into descending order and assigning a rank to each of their classes (folding patterns). The rank function $r_A(x)$ is then a function from D to $N = \{1, 2, ..., n\}$.

In order to properly compare and correctly combine score functions from multiple features, we have to normalize them. The normalization we used is the transformation from $s_A(x)$: $D \rightarrow R$ to $s_A^*(x)$: $D \rightarrow [0, 1]$ where $s_A^*(x) = \frac{s_A(x) - s_{\min}}{s_{\max} - s_{\min}}$, x in D and $s_{max} = max\{s_A(x) \mid x \text{ in } D\}$ and $s_{min} = min\{s_A(x) \mid x \text{ in } D\}$.

When m features are used to assign protein classes or folding patterns, there are 2^{m-1} combinations for all m individual features ($\sum_{k=1}^{m} \binom{m}{k} = 2^m - 1$) with rank or score functions. Hence, the total number of combinations to be considered for predicting protein class and protein folding pattern are $2^{m+1} - 2$ and $2^{2m+2} - 2^{m+3} + 4$ respectively in the HLA architecture. These numbers can become huge when the number of features m is large. It is very time-consuming to do all combinations for protein structure classification. Moreover, we have to evaluate the predictive power of each combination across all proteins. Therefore, in the current paper, we only consider combinations of two features which still retain fairly good prediction power. Combination of more than two features will be considered in our future work.

3.1 Methods of Combination and Feature Selection

Given m features A_i, $i = 1, 2, ..., m$, which assign score function s_{Ai} and rank function r_{Ai}, there are several different ways of combination. There are, among others, score combination, rank combination, voting,

average combination, and weighted combination [2, 3, 7-9, 11, 13, 17, 18]. In this paper, we use the average rank (or score) combination. For the m features A_i, rank functions r_{Ai}, and score functions s_{Ai}, we have the score function s_R and s_S of the rank combination and score combination respectively defined as:

$$s_R(x) = \sum_{i=1}^{m}[(r_{Ai}(x))/m], \text{ and } s_S(x) = \sum_{i=1}^{m}[(s_{Ai}(x))/m].$$

As we did before, $s_R(x)$ and $s_S(x)$ are then sorted into ascending and descending order to obtain the rank function of the rank combination $r_R(x)$ and the score combination $r_S(x)$, respectively.

Previous work in [2, 3, 7-9, 11, 13, 17, 18] have demonstrated that: (a) the combination of multiple features would improve the prediction accuracy only if (1) each of the feature functions has a relatively high performance, and (2) the individual features are distinctive (or diversified), and (b) rank combination performs better than score combination under certain conditions. In this paper, we use these rules (a)(1), (a)(2), and (b) as our guiding principle to select features and to decide on the method of combination [6]. A diversity function $d(A,B)$ between features A and B is then defined using the concept of the rank/score function defined by Hsu *et al*. [6, 8, 9].

3.2 The Diversity Rank/Score Graph

For each protein and feature A, we have the score function s_A and rank function r_A. As in IR [8, 9], we explore the scoring (and ranking) characteristics of feature A by calculating the rank/score function, $f_A : N \rightarrow [0, 1]$ as follows:

$$f_A(j) = (s_A^* \circ r_A^{-1})(j) = s_A^*(r_A^{-1}(j)).$$

We note that the set N is different from the set D which is the set of classes (or fold patterns). The set N is used as the index set for the rank function value. The rank/score function so defined signifies the scoring (or ranking) behavior of the feature A and is independent of the classes (or folding patterns) under consideration.

For protein p_i in $P = \{p_1, p_2, ..., p_t\}$ and the pair of features A and B, we define the **diversity score function** $d_i(A,B)$ as: $d_i(A,B) = \Sigma \mid f_A(j) - f_B(j) \mid$, where j is in $N = \{1, 2, \ldots, n\}$ and n is the number of classes (or folding patterns). When there are m' features selected, there are $\binom{m'}{2} = \frac{m'(m'-1)}{2}$ diversity score functions. If we let i vary and fix the feature pair (A,B), then $d_i(A,B)$ is the diversity score function $s_{(A,B)}(x)$ from $P = \{p_1, p_2, ..., p_t\}$ to R. Sorting $s_{(A,B)}(x)$ into descending order would lead to the **diversity rank function** $r_{(A,B)}(x)$. Consequently, the **diversity rank/score function** $f_{(A,B)}(x)$ is defined as:

$$f_{(A,B)}(j) = (s_{(A,B)} \circ r_{(A,B)}^{-1})(j) = s_{(A,B)}(r_{(A,B)}^{-1}(j)),$$

where j is in $T = \{1, 2, 3, ..., t\}$. Again we note that the set T is different from the set P which is the protein set considered. The set T is used as the index set for the diversity rank function value. The diversity rank/score function $f_{(A,B)}(k)$ so defined exhibits the diversity trend of the feature pair (A,B) and is independent of the specific protein under study.

The graph of the diversity rank/score function $f_{(A,B)}(j)$ is called the **diversity rank/score graph** (or **diversity graph** in short). In this paper, we aim to examine all the $\frac{m'(m'-1)}{2}$ diversity graphs to see which pair of features would give the highest diversity measurement. Following rules (a)(2) and (b), the rank combination of these two features is then calculated to give the final rank function and to choose the class (or folding pattern).

4 Results

The technique of combinatorial fusion [6] is used for protein structure classification on a testing data set with NN using RBFN under the HLA. Initially, we use eleven features, 'C' (reworded as A), 'CS' (as B), 'CSH' (as C), 'CSHP' (as D), 'CSHPV' (as E), 'CSHPVZ' (as F), 'CSHPVZ+B' (as G), 'CSHPVZ+B+SB' (as H), 'CSHPVZ+G' (as I), 'CSHPVZ+B+G' (as J), and 'CSHPVZ+B+SB+G' (as K), to assign protein classes for all proteins tested. Following the rule (a)(1), we select three features H, I, and K, for further combination because of their higher accuracy rate than others as demonstrated in [10]. With the help of rule (a)(1), we can reduce 2^{11}-1 combinations to 2^3-1 combinations. Following the rules (a)(2) and (b), we shall use the rank combination of the features to predict the protein class.

The diversity of any two of features H, I, and K is then calculated for all proteins tested and features H and I are found to have the highest diversity, as shown in Fig. 2, among all three feature combinations. Hence, we use the rank combination of features H and I to predict protein classes for all proteins tested. After the protein classes for all proteins tested have been predicted and categorized, the prediction of protein folding patterns follows in the HLA. We use the same rules and a diversity graph to choose a rank combination of features JK, HJ, HK, and HI to predict protein folding patterns in classes 1, 2, 3, and 4, respectively.

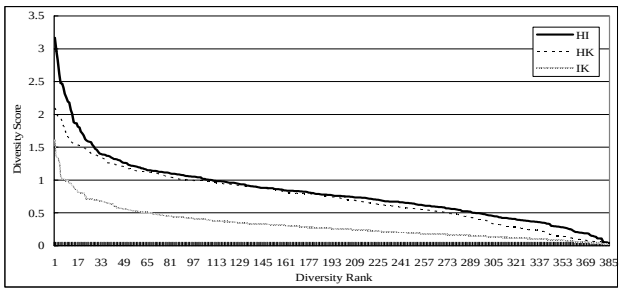


Fig. 2. The diversity rank/score graph for any pair of features (X,Y), X,Y in {H,I,K} for classifying protein classes

Table 2. The comparisons of overall prediction accuracy rates *Q* for protein classes

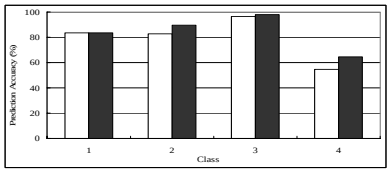
Method	HLA, 'CSHPVZ+B+SB'*, NN	HLA + data fusion, NN
Q	83.6	**87.8**

* Data from the paper (Huang *et al*. [10])

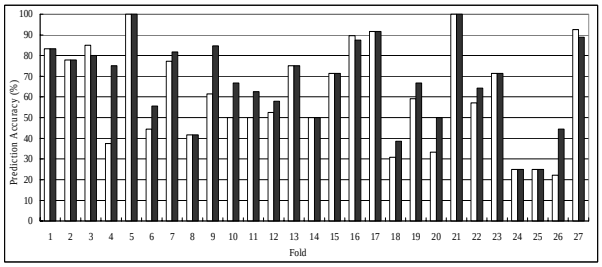
Table 3. The comparisons of overall prediction accuracy rates *Q* for protein folding patterns

Feature / Method	'CSH'	'CSHP'	'CSHPVZ'	'CSHPVZ +B+SB'
OvO[1], NN**	40.6	41.1	**41.8**	—
OvO[1], SVMs**	**45.2**	43.2	44.9	—
uOvO[2], SVMs**	**51.1**	49.4	49.6	—
AvA[3], SVMs**	56.0	**56.5**	53.9	—
HLA, NN*	53.3	54.3	56.4	**65.5**
HLA + data fusion, NN	**70.9**			

[1]one-versus-others method [5]; [2]unique one-versus-others method [5]; [3]all-versus-all method [5]
* Data from the paper (Huang *et al*. [10])
** Data from the paper (Ding and Dubchak [5])



(a) Protein classes



(b) Protein folds

Fig. 4. The comparisons of prediction accuracy rates Q_i of our previous work (Huang et al [10]) (in white) and the current work (in black) for 4 protein classes in (a) and 27 protein folding patterns in (b)

The standard percentage accuracy rate Q_i [6, 10, 14] is used to evaluate our work. $Q_i = p_i/n_i \times 100$, where n_i is the number of testing proteins in the *i*th class or folding pattern and p_i is the number of proteins being correctly predicted in the *i*th class or folding pattern. The overall prediction accuracy rate Q is given by $Q = \sum_{i=1}^{n} q_i Q_i$, where $q_i = n_i/K$, where K is the total number of proteins tested, and n is the number of classes or folding patterns. The overall prediction accuracy rates Q for protein classes in our previous [10] and current work are compared as shown in Table 2. The current overall prediction accuracy rate is 87.8%, 4.2% higher than that of our previous work. Table 3 shows that for prediction of folding pattern, our current work has an overall prediction accuracy rate of 70.9%, which is 14.4% higher than that of Ding and Dubchak [5], 5.4% higher than that of our previous work.

Fig. 4 shows the comparisons of Q_i of our previous work [10] and our current work. The current method gives Q_i ($\geq$80%) in 3 classes, especially in class α/β with accuracy rate reaches 97.9%, all higher than what we achieved previously, shown in Fig. 4(a). For protein folding patterns prediction, the current work gives Q_i ($\geq$80%) in 9 folding patterns, more than what in our previous work, as shown in Fig. 4(b). Also, the current work outperforms our previous work in 12 folding patterns, especially ($\geq$ 30% improvement) in folding patterns: α_4, β_3, β_4, $(\alpha/\beta)_5$, and $(\alpha+\beta)_2$. Hence, overall, there is an improvement with our current method.

5 Conclusions

Previous studied [6, 7-9, 13, 18] has been demonstrated that (a) the combination of multiple systems (or features) would improve the performance only if (1) each of the individual systems (features or functions) has a relatively high performance, and (2) each individual systems are distinctive (or different), and (b) combination by rank outperform combination by score under certain conditions.

In this paper, we use criterion (a)(1) to select features and then apply criterion (a)(2) by computing the diversity rank/score graph in order to select the pair of features with the highest diversity. Criterion (b) is then used to combine these two features using ranks. We apply combinatorial fusion in [6] to improve accuracy in protein structure prediction. We have successfully improved the overall predictive accuracy rate of 87.8% for the four classes and 70.9% for the 27 folding categories. We improve previous results by Huang *et al*. [10] and Ding and Dubchak [5]

by incorporating the method of combinatorial fusion in their approach using NN in HLA.

The work in this paper is one of a series of on-going projects towards the protein structure classification problem. In the previous one [12], we obtained the overall predictive accuracy rate of 87% for the four classes and 69.6% for the 27 folding categories by using combinatorial fusion for eight features without the gate function. The results in this paper show that the combinatorial fusion has potential to improve the prediction accuracy for protein structure classification by adding more features, even it has weak performance. This phenomenon encourages us to design new features for protein structure classification.

Acknowledgments

Post doctor fellowship of Chun Yuan Lin is supported by NSC under contract NSC93-3112-B-007-008. D. F. Hsu would like to thank National Tsing Hua University and the Ministry of Education for their support and hospitality during his visit to NTHU in Spring 2004.

Reference

[1] K.C. Chou and C.T. Zhang, Prediction of protein structural classes, *Crit. Rev. in Biochem. Mol. Biol.*, Vol.30, No.4, 1995, pp.275-349.

[2] H.Y. Chuang, H.F. Liu, S. Brown, C.M. Coffran, and D.F. Hsu, Identifying significant genes from microarray Data, *Proc. IEEE Symp. Bioinformatics and Bioengineering*, pp.358-365.

[3] H.Y. Chuang, H.F. Liu, F.A. Chen, C.Y. Kao, and D.F. Hsu, Combination methods in microarray analysis, *Proc. Intl. Symp. Parallel Architectures, Algorithms and Networks*, pp.625-630.

[4] L. Lo Conte, B. Ailey, T.J.P. Hubbard, S.E. Brenner, A.G. Murzin, and C. Chothia, SCOP: A structural classification of proteins database, *Nucleic Acids Res.*, Vol.28, No.1, 2000, pp.257-259.

[5] C.H.Q. Ding and I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks, *Bioinformatics*, Vol.17, No.4, 2001, pp.349-358.

[6] D.F. Hsu, Y.S. Chung, and B.S. Kristal, Combinatorial fusion analysis: method and practice of combining multiple scoring systems, to appear in *Advanced Data Mining Technologies in Bioinformatics*.

[7] D.F. Hsu and A. Palumbo, A study of data fusion in Cayley graphs G (Sn, Pn), *Proc. Intl. Symp. Parallel Architecture, Algorithms, and Networks*, pp.557-562.

[8] D.F. Hsu, J. Shapiro, and I. Taksa, Methods of data fusion in information retreival: rank vs. score combination, *DIMACS Technical Report* 58, 2002.

[9] D.F. Hsu and I. Taksa, Comparing rank and score combination methods for data fusion in information retrieval, *Information Retrieval*, Vol.8, 2005, pp.449-480.

[10] C.D. Huang, C.T. Lin, and N.R. Pal, Hierarchical Learning Architecture with Automatic Feature Selection for Multi-Class Protein Fold Classification, *IEEE Trans. NanoBioscience*, Vol.2, No.4, 2003, pp.503-517.

[11] M.A. Kuriakose, W.T. Chen, Z.M. He, A.G. Sikora, P. Zhang, Z.Y. Zhang, W.L. Qiu, D.F. Hsu, C.M. Coffran, S.M. Brown, E.M. Elango, M.D. Delacure, and F.A. Chen, Selection and Validation of Differentially Expressed Genes in Head and Neck Cancer, *Cellular and Mol. Life Sci.*, Vol.61, 2004, pp.1372-1383.

[12] C.Y. Lin, K.L. Lin, C.D. Huang, H.M. Chang, C.Y. Yang, C.T. Lin, C.Y. Tang, D.F. Hsu, "Feature Selection and Combination Criteria for Improving Predictive Accuracy in Protein Structure Classification," to appear in *IEEE Symp. Bioinformatics and Bioengineering* 2005.

[13] K.B. Ng and P.B. Kantor, Predicting the effectiveness of naïve data fusion on the basis of system characteristics, *J. American Society for Information Sci.*, Vol.51. No.13, 2000, pp.1177-1189.

[14] B. Rost and C. Sander, Prediction of protein secondary structure at better than 70% accuracy, *J. Mol. Biol.*, Vol.232, 1993, pp.584-599.

[15] V.N. Vapnik, *The Nature of Statistical Learning Theory*, New York: Springer-Verlag, 1995.

[16] C.H. Wu, *Neural Networks and Genome Informatics*. Amsterdam, The Netherlands: Elsevier, 2000.

[17] L. Xu, A. Krzyzak, and C.Y. Suen, Method of Combining Multiple Classifiers and their Application to Handwriting Recognition, *IEEE Trans. SMC*, Vol.22, 1992, pp.418-435.

[18] J.M. Yang, Y.F. Chen, T.W. Shen, B.S. Kristal, and D.F. Hsu, Consensus scoring criteria for improving enrichment in virtual screening, *Journal of Chemical Information and Modeling*, 2005, pp.1134-1146.