BioChem Press   http://www.biochempress.com

Inter*net* Electronic Journal of
Molecular Design

Support Vector Machines for Predicting Protein Homo–Oligomers by Incorporating Pseudo–Amino Acid Composition

Shao–Wu Zhang, Quan Pan, Hong–Cai Zhang, Yong–Hong Wu, and Jian–Yu Shi

Department of Automatic Control, Northwestern Polytechnic University, Xi'an 710072, P. R. China

Citation of the article:
S.–W. Zhang, Q. Pan, H.–C. Zhang, Y.–H. Wu, and J.–Y. Shi, Support Vector Machines for Predicting Protein Homo–Oligomers by Incorporating Pseudo–Amino Acid Composition, *Internet Electron. J. Mol. Des.* **2003**, *2*, 392–402, http://www.biochempress.com.

Support Vector Machines for Predicting Protein Homo–Oligomers by Incorporating Pseudo–Amino Acid Composition[#]

Shao–Wu Zhang,* Quan Pan, Hong–Cai Zhang, Yong–Hong Wu, and Jian–Yu Shi

Department of Automatic Control, Northwestern Polytechnic University, Xi'an 710072, P. R. China

Abstract
Following the success of human genome project, the gap between sharply increasing the number of protein sequences entering into data bank and slow accumulation of know structure is becoming large. Developing a fast and accurate method to predict the protein properties based on the primary sequences becomes indispensable. In general, the performance of the predictive system can be improved by selecting appropriate algorithm and the fitting method of extracting feature. Thus a new method of extracting feature (the weighting pseudo–amino acid composition) from the sequences has been introduced to predict the protein homo–oligomers, which is a combination of a set of weighting discrete sequence correlation factors computed with the amino acid index profile and the 20 components of the conventional amino acid composition. We extract four attribute parameter datasets (COMP, PLIV, FAUJ and MAXF) from the primary sequences as examples to investigate this problem. The COMP attribute dataset is composed of amino acid composition, and the PLIV, FAUJ and MAXF attribute datasets are composed of the amino acid composition and a set of weighting discrete sequence correlation factors of corresponding amino acid residue index. The total accuracies of PLIV, FAUJ and MAXF using support vector machines (SVM) algorithm are 80.36%, 79.34% and 79.02% respectively in 10 fold cross–validation (10CV) test, which are 4.59%, 3.57% and 3.25% respectively higher than that of COMP. Based on the same COMP and PLIV attribute datasets, the total accuracies of SVM are 33.87% and 18.05% respectively higher than that of covariant discriminant algorithm in the jackknife test. These results show that the method of extracting feature from the protein sequences is effective and feasible for predicting homo–oligomers, and implies that the primary sequences of homo–oligomeric proteins contain quaternary structure information, and also indicates that the performance of SVM is superior to the covariant discriminant algorithm for classifying protein homo–oligomers.

Keywords. Support vector machines; SVM; covariant discriminant; weighting pseudo–amino acid composition; amino acid composition; homo–oligomers.

1 INTRODUCTION

The functional diversity of proteins is made possible by the diversity of their spatial structures, which are capable of highly specific molecular recognition. Understanding or simulating the molecular processes involved in the formation of protein structure and in their biological function is

---

a major challenge of molecular biology. Although the protein spatial structure can be determined by conducting various experiments, it is time consuming and costly to acquire this kind of knowledge solely by experiments [1]. Nowadays, it is generally accepted that the protein structure is determined by its amino acid sequence [2,3], and the number of protein sequences entering into databanks has been rapidly increasing, thus, predicting the spatial structure based on a given protein primary sequence information could play a significant role, in conjunction with experimental methods.

Some proteins consist of more than one polypeptide chain or subunit. These are also called multimeric proteins, and may be formed by several identical polypeptides or by different ones. Quaternary structure of a protein is refers to the class, number, spatial arrangement of subunits and interaction of non–covalently bound monomeric protein subunits to form oligomers. Such complexes are common in eukaryotic cells and are involved in many critical cellular processes, such as metabolism, cell signaling and chromosome replicating etc. Many previous studies are devoted to the analysis of the protein–protein interactions and the prediction of interaction sites from the known 3D structures and sequence profile [4–10]. Robert Garian studied the predicting of homodimers and non–homodimers using decision–tree models and got the result that protein primary sequence contains quaternary structure information [11]. The purpose of this study is to develop a reliable prediction system of homo–oligomers by introducing a new method of extracting feature from the protein sequences, the weighting pseudo–amino acid composition, and Vapnik's Support Machines [12,13] to discriminate the homodimers, homotrimers, homotetramers and homohexamers.

2 MATERIALS AND METHODS

2.1 Database

The database R was selected from Robert Garian's database [11]. It was consisted of 1568 homo–oligomeric protein sequences, 914 of which were homodimers (2EM), 139 homotrimers (3EM), 407 homotetramers (4EM) and 108 homohexamers (6EM). The database was obtained from Release 34 of the SWISS–PROT database [14] and limited to the prokaryotic, cytosolic subset of homo–oligomers in order to eliminate membrane proteins and other specialized proteins.

2.2 Support Vector Machines

Support Vector Machines (SVM) is a new type of learning machines based on statistical learning theory, which is currently considered as one of the most efficient method in many real–world applications. Due to SVM powerful classification, it was applied with success in medicine, computational biology, and structure–activity relationships, including microarray gene expression data [15], translation initiation sites [16], protein class [17], membrane protein type [18], protein–

protein interactions [10], aquatic toxicity mechanism [19], carcinogenic activity [20], structure–odor relationship [21], protein subcellular localization [22–24], and protein fold [25].

The SVM works by mapping the data samples, which are points in the input space, into a higher–dimensional space called the feature space. An optimal separating hyperplane (OSH) can then be defined in the feature space. The mapping function used only involves the relatively low–dimensional vectors in the input space and dot products in the feature space. This dot product is represented by a kernel function in the input space. The separating hyperplane can be determined without having to represent the feature space explicitly. What follows is a brief description of the SVM algorithm. A more detailed description can be found in Vapnik's book [13] and Cristianini's book [26].

For a two–class classification problem, assume that we have a set of samples, *i.e.* a series of input vectors $\vec{x}_i \in R^d$ ($i = 1, 2,…, N$), with corresponding labels $y_i \in \{+1,-1\}$ ($i = 1, 2, …, N$). Here, +1 and −1 indicate the two classes. The goal here is to construct one binary classifier or derive one decision function from the available samples, which has small probability of misclassifying a future sample. Both the basic linear separable case and the most useful linear non–separable case (for most real life problems) are considered here.

For a linear separable case, there exists a separating hyperplane whose function is $\vec{w} \bullet \vec{x} + b = 0$, which implies the following:

$$y_i(\vec{w} \bullet \vec{x}_i + b) \geq 1 \quad i=1, 2, …, N$$

By minimizing $\frac{1}{2}\|\vec{w}\|^2$ subject to this constraint, the SVM approach tries to find a unique separating hyperplane, which maximizes the distance between the hyperplane and the nearest data points of each class. The classifier is called the largest margin classifier.

By introducing Lagrange multipliers α_i, the SVM training procedure amounts to solving a convex quadratic programming (QP) problem. The solution is a unique globally optimized result, and can be shown as the following formula:

$$\vec{w} = \sum_{i=1}^{N} y_i \alpha_i \vec{x}_i$$

Only if the corresponding $\alpha_i > 0$, these x_i are called Support Vectors. When a SVM is trained, the decision function can be written as:

$$f(\vec{x}) = \text{sgn}(\sum_{i=1}^{N} y_i \alpha_i \vec{x} \bullet \vec{x}_i + b)$$

For a linear non–separable case, in order to allow for training errors, this can be done by introducing positive slack variables ζ_i ($i = 1, 2,…, N$) in the constraints, which then become:

$$y_i(\vec{w} \bullet \vec{x}_i + b) \geq 1 - \xi_i \quad \xi_i \geq 0, \forall i$$

We want to simultaneously maximize the margin and minimize the number of misclassifications. This can be achieved by changing the objective function from $\frac{1}{2}\|\vec{w}\|^2$ to

$$\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N}\xi_i^k$$

$$\text{Minimize} \quad \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N}\xi_i^k$$

$$\text{Subject to} \quad y_i(\vec{w} \bullet \vec{x}_i + b) - 1 + \xi_i \geq 0 \ , \ i = 1,2,\cdots,N$$

$$\xi_i \geq 0 \quad i = 1,2,\cdots,N$$

The error weight C is a regularization parameter to be chosen by the user, it means the size of penalties assigned to errors. The optimization problem is convex for any positive integer k, for $k = 1$ and $k = 2$ it is also a quadratic programming problem. This is called the Soft Margin Generalization of the OSH, while the original concept with no errors allowed is called Hard Margin.

For a two–class nonlinear classification problem, SVM performs a nonlinear mapping $\Phi(\bullet)$ of the input vector $\vec{x}$ from the input space R^d into a higher dimensional Hilbert space H, and constructs an Optimal Separating Hyperplane. In the linear separable case, we know that the algorithm only depends on inner products between training examples and test examples. So we can generalize to nonlinear case. The inner products are substituted by the kernel function $k(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i) \bullet \Phi(\vec{x}_j)$ in the input space. Then, the decision function implemented by SVM can be written as:

$$f(\vec{x}) = \text{sgn}(\sum_{i=1}^{N} y_i \alpha_i k(\vec{x}, \vec{x}_i) + b)$$

Two typical kernel functions are listed below;

$$\text{Polynomial function} \quad k(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \bullet \vec{x}_j + 1)^d$$

$$\text{Radial basis function (RBF)} \quad k(\vec{x}_i, \vec{x}_j) = \exp(-\gamma\|\vec{x}_i - \vec{x}_j\|^2)$$

The software used to implement SVM was SVM[light] by Joachims [27] which can be freely downloaded from http://ais.gmd.de/~thorsten/svm_light/ for academic use. The core optimization method for solving the QP problem was based on the 'LOQO' algorithm [28].

2.3 Extracting the Sequence Descriptor

According to the studies of Nakashima [29], Klein [30] and Chou [31,32], the 20–D (dimension) attribute vector is used to represent a protein primary sequence, which defined as:

$$\vec{x} = [f_1, f_2, \cdots, f_{20}]^T$$

where f_i (i = 1, 2,…, 20) are the occurrence frequencies of the 20 amino acids in the protein

concerned, arranged alphabetically according to their signal letter codes, and *T* means a transpose operator.

Since the information within the primary sequence is greatly reduced by considering the amino acid composition alone, the sequence order of amino acids in the query protein should be taken into account. Thus a set of weighting sequences order–correlated factors based on the physicochemical properties of amino acid along the primary sequence of the query protein have been considered. In other words, in addition to the 20–D components of the amino acid frequencies, other *m*–D components should be added to form a (20 + *m*)–D vector. The attribute vector is defined as:

$$\vec{x} = \left[f_1, f_2, \cdots, f_{20}, r_1, r_2, \cdots, r_m \right]^T$$

where f_i (*i* = 1, 2, …, 20) are the occurrence frequencies of the 20 amino acids in the protein concerned, r_n (*n* = 1, 2,…, *m*) are the weighting sequences order–correlated factors, and *m* is an integer to be determined by the optimum prediction, and the (20 + *m*)–D vector was called pseudo–amino acid composition vector [1]. The calculation of weighting sequences order–correlated factors will be shown as follows.

Consider a protein chain of *L* amino acid residues: $R_1R_2R_3R_4R_5R_6…R_L$. The weighting sequence order–correlated factors r_n are defined as:

$$r_n = \frac{w}{L-n} \sum_{i=1}^{L-n} [H(R_{i+n}) - H(R_i)]^2 \qquad n=1, 2, \ldots, m \tag{1}$$

where $H(R_{i+n})$, $H(R_i)$ are the index values of amino acid R_{i+n} and R_i respectively, and *w* is weight factor. The index values of amino acid R_{i+n} and R_i can be selected from Kanehisa's Amino Acid Index database [33], which may be accessed through the DBGET/LinkDB system at GenomeNet (http://www.genome.ad.jp/dbget) or may be downloaded by anonymous FTP (ftp://genome.ad.jp/db/genoment/aaindex). An amino acid index is a set of 20 numerical values representing any of the different physicochemical properties of the 20 amino acids.

Table 1. Four datasets extracted from protein sequences

Symbol	Attribute parameter illustration
COMP	Amino acids composition
PLIV [1]	Pseudo–amino acid composition extracting from amino acid residue index of Pliska *et al.*
FAUJ [2]	Pseudo–amino acid composition extracting from amino acid residue index of Fauchere *et al.*
MAXF [3]	Pseudo–amino acid composition extracting from amino acid residue index of Maxfield–Scheraga

[1] PLIV810101 partition coefficient (Pliska *et al.*, 1981); [2] FAUJ880103 normalized van der Waals volume (Fauchere *et al.*, 1988); [3] MAXF760102 normalized frequency of extended structure (Maxfield–Scheraga, 1976)

According to the amino acid frequencies and the weighting sequences order–correlated factors, we extracted four attribute parameter sets from the primary sequences, which are clearly shown in Table 1. As an example, protein FER_DESGI can be computed and represented by a vector in 43–D space: [10.34 10.34 13.79 15.52 1.72 1.72 0.00 8.62 1.72 1.72 3.45 5.17 6.90 0.00 1.72 5.17 0.00

12.07 0.00 0.00 7.05 6.94 7.97 7.16 7.77 7.53 6.09 6.96 6.31 6.46 7.51 7.21 5.73 8.34 6.99 7.66 8.46 7.60 6.03 7.35 6.75 6.05 4.93]^T in PLIV dataset, here, $w = 10$, $m = 23$.

2.4 Design and Implementation of the Prediction System

Protein homo–oligomers prediction is a multi–class classification problem. Here, the class number is equal to 4. A simple strategy to handle the multi–class classification is to reduce the multi–classification to a series of binary classifications. For a k–class classification, k SVMs are constructed. The ith SVM will be trained with all of the samples in the ith class with positive labels and all other samples with negative labels. We refer to SVMs trained in this way as 1–v–r SVMs (short for one–versus–rest). Finally one unknown sample is classified into the class that corresponds to the 1–v–r SVM with the highest output value. This method was used to construct a prediction system (*i.e.* one 4–class classifier) for protein homo–oligomers.

2.5 Classification of System Assessment

The classification quality can be examined using the jackknife test and 10–fold cross–validation (10CV) test, which are objective and rigorous testing procedures. The total prediction accuracy (Q), the prediction accuracy (Q_i) and Matthew's Correlation Coefficient ($MCC(i)$) [34] for each class of homo–oligomers calculated for assessment of the prediction system are given by:

$$Q = \sum_{i=1}^{l} p(i)/N \quad Q_i = p(i)/obs(i) \quad MCC(i) = \frac{p(i)n(i)-u(i)o(i)}{\sqrt{(p(i)+u(i))(p(i)+o(i))(n(i)+u(i))(n(i)+o(i))}}$$

where N is the total number of sequences, l is the class number, $obs(i)$ is the number of sequences observed in i class protein homo–oligomers, $p(i)$ is the number of correctly predicted sequences of i class protein homo–oligomers, $n(i)$ is the number of correctly predicted sequences not of i class protein homo–oligomers, $u(i)$ is the number of under–predicted sequences of i class protein homo–oligomers and $o(i)$ is the number of over–predicted sequences of i class protein homo–oligomers.

3 RESULTS AND DISCUSSION

3.1 The results with SVM in the 10CV Test

From table 2, we can see that the total accuracy of COMP based only on amino acid composition in 10CV test is 75.77%, and the total accuracy for PLIV, FAUJ and MAXF based on amino acid composition adding a set of weighting sequences order–correlated factors are 80.36%, 79.34%, 79.02% respectively, which are 4.59%, 3.57%, 3.25% respectively higher than that of COMP set. The MCC of each class (2EM, 3EM, 4EM and 6EM) for the PLIV, FAUJ and MAXF is bigger than that of the corresponding class for COMP. These results indicate that the method of extracting feature from the protein sequences is effective and feasible, and the SVM can be applied to predict the protein homo–oligomers.

Table 2. The predictive results using RBF kernel function support vector machines ($C = 1000$) in 10CV test

	COMP ($\gamma = 0.04$)		PLIV ($\gamma = 0.03$, $w = 10$, $m = 23$)		FAUJ ($\gamma = 0.03$, $w = 1$, $m = 29$)		MAXF ($\gamma = 0.03$, $w = 100$, $m = 22$)	
	Accuracy %	MCC	Accuracy %	MCC	Accuracy %	MCC	Accuracy %	MCC
2EM	89.28	0.5475	93.00	0.6210	92.23	0.6041	93.00	0.6005
3EM	51.08	0.6565	58.27	0.7431	55.40	0.7028	53.24	0.6975
4EM	62.65	0.5446	68.06	0.6535	65.85	0.6204	65.85	0.6267
6EM	42.59	0.5550	48.15	0.6168	51.85	0.6625	43.52	0.5977
Total accuracy %	75.77	–	80.36	–	79.34	–	79.02	–

We have analyzed 472 sets of indices in AAindex Ver5.0. The total accuracy in 10CV test is used to evaluate the predictive ability of each amino acid index. Among 472 sets of data, about 80% could differently improve the classifying results. By the hierarchical clustering [35], the 472 indices can be divided into six major classes: α and turn propensities, β propensity, amino acid composition, hydrophobicity, physicochemical properties, and other properties. We found that most of hydrophobicity amino acid indices used for classifying have better performance than that of other five classes of amino acid index, suggesting that biologically relevant complex formation is driven predominantly by the hydrophobic effect [7]. The results listed in Table 2 are three typical examples of several indices. The amino acid indices of PLIV, FAUJ and MAXF belong to the class of hydrophobicity, physicochemical properties and β propensity respectively, which have the best performance in their each class. These results also imply that hydrophobic interactions are the greatest contribution to subunit interaction, hydrogen bonds and van der Waals interactions also contribute to the specificity of subunit interaction. In addition, we try to add up several different amino acid indices according to Chou's method [1], but the classifying results have not been improved apparently. We think that adding up several different amino acid indices may exist the problem of information fusion.
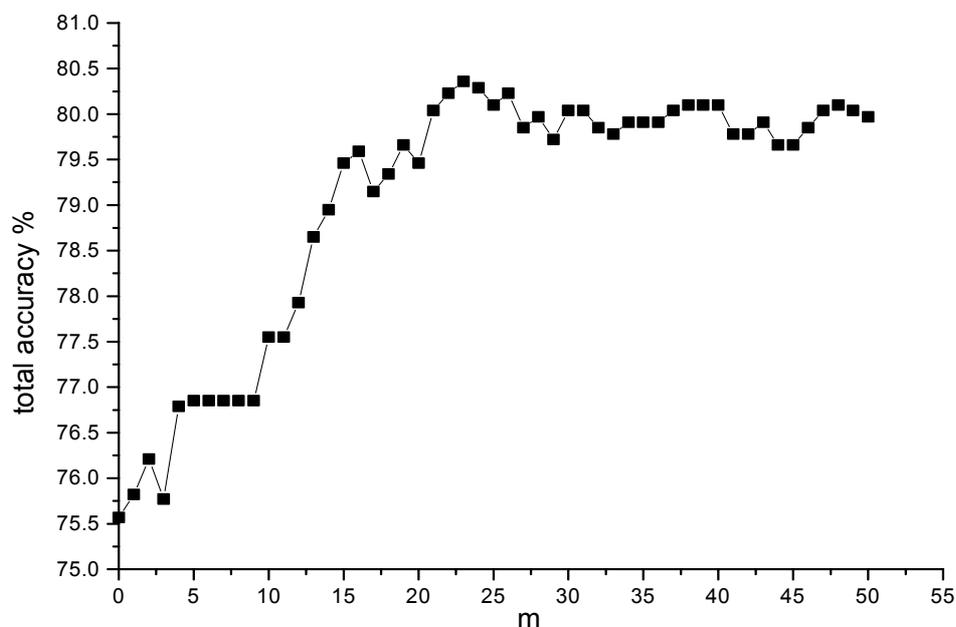

Figure 1. The relationship between m used in the prediction (x–axis) and the total predictive accuracy (y–axis) in the 10Cv test. The highest accuracy is achieved at $m = 23$.

The number of the weighting sequence order–correlated factors used in PLIV, FAUJ and MAXF parameter sets is denoted by *m* in Eq. (1). The classifying results of the PLIV, FAUJ and MAXF parameter sets can be affected with different *m* values, thus we take PLIV set as an example to study the classifying result affected with different *m* values in the same condition *w* = 10. The results are clearly shown in Figure 1, were the *x*–axis represents the number of the weighting sequence order–correlated factors used in the prediction, whereas the *y*–axis represents the total accuracy in 10CV test. Obviously, there is an optimal value of *m*, for example, when *m* = 23, the best total accuracy 80.36% can be obtained.

3.2 Comparison with the Covariant Discriminant Algorithm

The SVM method predictions were compared with the Covariant Discriminant algorithm [1,36,37] in a jackknife test. The results are shown in the Table 3.

Table 3. The comparison of the RBF kernel function SVM and the Covariant Discriminant algorithm in jackknife test

	SVM (*C*=1000)				Covariant Discriminant			
	COMP (γ = 0.04)		PLIV (γ = 0.03, *w* = 10, *m* = 23)		COMP		PLIV	
	Accuracy %	MCC	Accuracy %	MCC	Accuracy %	MCC	Accuracy %	MCC
2EM	89.93	0.5825	93.22	0.6368	34.68	0.2663	59.74	0.3810
3EM	57.55	0.6911	62.59	0.7623	59.71	0.4582	59.71	0.6470
4EM	64.13	0.5715	68.55	0.6599	51.11	0.2489	77.15	0.4061
6EM	46.30	0.5855	54.63	0.6829	68.52	0.2167	47.22	0.4663
Total accuracy %	77.36	–	81.44	–	43.49	–	63.39	–

Table 3 shows that the performance of the SVM method is superior to the covariant distriminant algorithm. In addition, the results of the PLIV are always better than that of COMP in both of algorithms. These results show that the weighting pseudo–amino acid composition may include some order information of the protein homo–oligomers.

3.3 The Performance of the Predictive System Influenced by the Size of Database and the Unbalance of Sample Numbers Between the Four Classes

To investigate the influence of the database size and the sample unbalance between the four classes, we established subset Database A. The Database A is randomly selected from the Database R, which consists of 432 homo–oligomeric protein sequences. Each class (2EM, 3EM, 4EM and 6EM) has 108 protein sequences in the Database A. The results are shown in the Table 4. The results of the Database A are the mean of five random selections. It is seen that the database size and the sample unbalance between classes have great influence to the performance of the predictive system. Generally, increasing the number of the training set and decreasing the unbalance of the samples between classes can improve the performance of the predictive system, and enhance the system stability. In addition, we should see that the performance of PLIV is still better than that of COMP in Database A. This result demonstrates again that the weighting pseudo–amino acid

composition may present useful information and hence improve the prediction with properly joining the amino acid composition.

Table 4. The performance of the predictive system influenced by the database size and the sample unbalance between the classes using RBF kernel function support vector machines ($C = 1000$) in the jackknife test

	Database R				Database A			
	COMP		PLIV		COMP		PLIV	
	Accuracy %	MCC	Accuracy %	MCC	Accuracy %	MCC	Accuracy %	MCC
2EM	89.93	0.5825	93.22	0.6368	48.52	0.3360	53.33	0.4014
3EM	57.55	0.6911	62.59	0.7623	75.93	0.5948	77.41	0.6783
4EM	64.13	0.5715	68.55	0.6599	57.59	0.4725	63.89	0.5377
6EM	46.30	0.5855	54.63	0.6829	65.00	0.5537	74.07	0.6275
Total accuracy %	77.36	–	81.44	–	61.76	–	67.18	–

3.4 SVM Parameters Selection

SVM still has a few adjustable parameters that need to be determined. SVM training includes the selection of the proper kernel function and their parameters. Both of polynomial kernel and RBF kernel were selected to study, because the successful theoretical methods are not available for kernel function types and parameters selection. By studying, we found that the regularization parameter C had little influence on the classifying performance for two types of kernel function, so we selected the default $C = 1000$ of SVM[light] program. For polynomial kernel, the algorithm is divergence or the training speed is very slow, thus we did not select it for classification. The parameter γ of RBF kernel has different effects on classification performance. Thus, we can select the best kernel types and parameters by computer operation for different datasets.

4 CONCLUSIONS

The results of computation experiments have shown that the method of extracting feature by incorporating the weighting pseudo–amino acid composition is effective and feasible, and the SVM can be applied to predict the homo–oligomers from the protein sequences. The feature vectors composed of amino acid composition and pseudo–amino acid composition may contain protein quaternary structure information, and appear to capture essential information about the composition and hydrophobicity of residues in the surface patches that buried in the interfaces of associated subunits. Although these feature vectors can reflect protein quaternary structure information at a certain extent, but these methods of representing protein sequence have a certain limitation. Due to many amino acid indices and the selectivity of weight factor w and m values of sequences order–correlated factors, there are many forms of amino acid composition integrating with sequences order–correlated factors. Thus, the best classifying result can be obtained for a given dataset by optimal selecting amino acid index, m value and weight factor w.

Acknowledgment

5 REFERENCES

[1] K. C. Chou, Prediction of Protein Cellular Attributes Using Pseudo–Amino Acid Composition, *Proteins*: *Struct*. *Funct*. *Genet*. **2001**, *43*, 246–255.

[2] C. B. Anfinsen, E. Haber, M. Sela and F. H. White, The Kinetics of Formation of Native Ribonuclease During Oxidation of the Reduced Polypeptide Chain, *Proc*. *Natl*. *Acad*. *Sci*. *USA*. **1961**, *47*, 1309–1314.

[3] C. B. Anfinsen, Principles that Govern the Folding of Protein Chains, *Science* **1973**, *181*, 223–230.

[4] Y. Ofran and B. Rost, Analysing Six Types of Protein–Protein Interfaces, *J*. *Mol*. *Biol*. **2003**, *325*, 377–387.

[5] I. M. A. Nooren and J. M. Thornton, Structural Characterisation and Functional Significance of Transient Protein–Protein Interactions, *J*. *Mol*. *Biol*. **2003**, *325*, 991–1018.

[6] H. X. Zhou and Y. Shan, Prediction of Protein Interaction Sites from Sequence Profile and Residue Neighbor List, *Proteins*. **2001**, *44*, 336–343.

[7] F. Glaser, D. M. Steinberg, I. A. Vakser and N. Ben–Tal, Residue Frequencies and Pairing Preference at Protein–Protein Interfaces, *Proteins*: *Struct*. *Funct*. *Genet*. **2001**, *43*, 89–102.

[8] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates and D. Eisenberg, Detecting Protein Function and Protein–Protein Interactions from Genome Sequences, *Science*. **1999**, *285*, 751–753.

[9] S. Jones and J. M. Thornton, Analysis of Protein–Protein Interaction Sites Using Surface Patches, *J*. *Mol*. *Biol*. 1997, *272*, 121–132.

[10] J. R. Bock and D. A. Gough, Predicting Protein–Protein Interactions from Primary Structure, *Bioinformatics* **2001** *17*, 455–460.

[11] R. Garian, Prediction of Quaternary Structure from Primary Structure, *Bioinformatics*. **2001**, 17, 551–556.

[12] V. Vapnik (Ed.), *The Nature of Statistical learning Theory*, Springer, New York, 1995.

[13] V. Vapnik (Ed.), *Statistical Learning Theory*, Wiley, New York, 1998.

[14] A. Bairoch and R. Apweiler, The SWISS–PROT Protein Data Bank and Its New Supplement TrEMBL, *Nucleic Acids Res*. **1996**, *24*, 21–25

[15] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr. and D. Haussler, Knowledge–based Analysis of Microarray Gene Expression Data by Using Support Vector Machines, *Proc*. *Natl*. *Acad*. *Sci*. *USA*. **2000**, *97*, 262–267.

[16] A. Zien, G. Ratsch, S. Mika, B. Schölkopf, T. Lengauer and K. R. Muller, Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites, *Bioinformatics* **2000**, *16*, 799–807.

[17] Y. D. Cai, X. J. Liu, X. B. Xu and K. C. Chou, Prediction of Protein Structural Classes by Support Vector Machines, *Comput*. *Chem*. **2002**, *26*, 293–296.

[18] Y.–D. Cai, X.–J. Liu, X. Xu, and K.–C. Chou, Support Vector Machines for Predicting Membrane Protein Types by Incorporating Quasi–Sequence–Order Effect, *Internet Electron*. *J*. *Mol*. *Des*. **2002**, *1*, 219–226, http://www.biochempress.com.

[19] O. Ivanciuc, Support Vector Machine Identification of the Aquatic Toxicity Mechanism of Organic Compounds, *Internet Electron*. *J*. *Mol*. *Des*. **2002**, *1*, 157–172, http://www.biochempress.com.

[20] O. Ivanciuc, Support Vector Machine Classification of the Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons, *Internet Electron*. *J*. *Mol*. *Des*. **2002**, *1*, 203–218, http://www.biochempress.com.

[21] O. Ivanciuc, Structure–Odor Relationships for Pyrazines with Support Vector Machines, *Internet Electron*. *J*. *Mol*. *Des*. **2002**, *1*, 269–284, http://www.biochempress.com.

[22] S. J. Hua and Z. R. Sun, Support Vector Machine Approach for protein Subcellular Localization Prediction, *Bioinformatics* **2001**, *17*, 721–728.

[23] Y. D. Cai, X. J. Liu, X.B. Xu and K.C. Chou, Support Vector Machines for Prediction of Protein Subcellular Location, *Mol*. *Cell Biol*. *Res*. *Commun*. **2000**, *4*, 230–233.

[24] Y. D. Cai, X. J. Liu, X.B. Xu and K.C. Chou, Support Vector Machines for Prediction of Protein Subcellular Location by Incorporating Quasi–Sequence–Order Effect, *J*. *Cell*. *Biochem*. **2002**, *84*, 343–348.

[25] C. H. Q. Ding and I. Dubchak, Multi–Class Protein Fold Recognition Using Support Vector Machines and Neural Networks, *Bioinformatics* **2001**, *17*, 349–358.

[26] N. Cristianini and J. Shawe–Taylor (Ed.), *An Introduction to Support Vector Machines and Other Kernel–Based Learning Methods*, Cambridge University Press, 2000.

[27] T. Joachims, Making large–scale SVM learning practical; in: *Advances in Kernel Methods–Support Vector learning*, Eds. B. Scholkopf, C. Burges and A. Smola, MIT Press, Cambridge, MA, 1999, pp.42–56.

http://www.biochempress.com

[28] R. J. Vanderbei, Interior Point Methods: Algorithms and Formulations, *ORSA J. Comput*. **1994**, *6*, 32–34.

[29] H. Nakashima, K. Nishikawa and T. Ooi, The Folding Type of A Protein is Relevant to the Amino Acid Composition, *J. Biochem*. **1986**, *99*, 152–162.

[30] P. Klein, Prediction of Protein Structural Class by Discriminat Analysis, *Biochem. Biophys. Acta*. **1986**, *876*,205–275.

[31] K. C. Chou and G. M. Maggiora, Domain Structural Prediction, *Protein Eng*. **1998**, *11*, 523–538.

[32] K. C. Chou, A Key Driving Force in Determination of Protein Structural Classes, *Biochem. Biophys. Res. Commun*. **1999**, *264*, 216–224.

[33] S. Kawashima, H. Ogata and M. Kanehisa, AAindex: Amino Acid Index Database. *Nucleic Acids Res*. 1999, *27*, 368–369.

[34] G. D. Fasman (Ed), *Handbook of Biochemistry and Molecular Biology*, 3rd ed., Proteins–Volume1, CRC Press, Cleveland, 1976.

[35] K. Tomii and M. Kanehisa, Analysis of Amino Acid Indices and Mutation Matrices for Sequence Comparison and Structure Prediction of Protein, *Protein Eng*. **1996**, *9*, 27–36.

[36] R. O. Duba and P. E. Hart (Ed.), *Pattern Classification and Scene Analysis*, Chap. 2, John Wiley & Sons, New York, 1973.

[37] K. C. Chou and D. W. Elord, Protein Subcellular Location Prediction, *Protein Eng*. **1999**, *12*, 107–108.