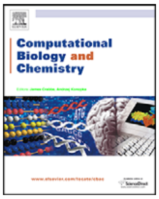
Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem



Algorithm Note

Variable predictive model based classification algorithm for effective separation of protein structural classes

Rao Raghuraj, S. Lakshminarayanan*

Department of Chemical and Biomolecular Engineering, National University of Singapore,
4 Engineering Drive 4, Singapore 117576, Singapore

ARTICLE INFO

Article history:
Received 2 March 2007
Received in revised form 22 March 2008
Accepted 24 March 2008

Keywords:
Protein classification
Secondary structure prediction
Variable predictive model
Discriminant analysis
Computational biology

ABSTRACT

Variable predictive model based class discrimination (VPMCD) algorithm is proposed as an effective protein secondary structure classification tool. The algorithm mathematically represents the characteristics amino acid interactions specific to each protein structure and exploits them further to distinguish different structures. The new concept and the VPMCD classifier are established using well-studied datasets containing four protein classes as benchmark. The protein samples selected from SCOP and PDB databases with varying homology (25–100%) and non-uniform distribution of class samples provide challenging classification problem. The performance of the new method is compared with advanced classification algorithms like component coupled, SVM and neural networks. VPMCD provides superior performance for high homology datasets. 100% classification is achieved for self-consistency test and an improvement of 5% prediction accuracy is obtained during Jackknife test. The sensitivity of the new algorithm is investigated by varying model structures/types and sequence homology. Simpler to implement VPMCD algorithm is observed to be a robust classification technique and shows potential for effective extensions to other clinical diagnosis and data mining applications in biological systems.

1. Introduction

Prediction of physical structures and subsequent separation into characteristic groups is important for analyzing the functional influences of biologically vital proteins. To analyze these structures, proteins are usually classified into one of the four secondary structure classes: helices (α), strands (β) and different composition of both of them (α/β and $\alpha+\beta$) (Zhou, 1998). Thousands of experimentally established proteins available in public databases with known structures, if *learnt* accurately, can be used to predict the structure of an unknown protein. Such an attempt using machine learning and computational algorithms is crucial for avoiding expensive and time-consuming experimental predictions, especially given the possible existence of huge sets of proteins with unknown structures/functions in nature. The overall protein structure classification problem can be always formulated as a multivariable, multi-class discriminant analysis problem. Classification based on various inter-class distance measures (LDA/QDA) (Nakashima et al., 1986; Edler et al., 2001), component coupled discrimination (CCD) (Chou, 1995), neural network approach (ANN) (Cai and Zhou, 2000), support vector machines (SVMs) (Cai et al., 2001) and rough sets (Cao et al., 2006) have all been attempted to infer the protein secondary structures with varying degree of success. The component coupled algorithm, SVM and ANN methods are the widely accepted classification methods and have been successfully employed for most of the protein classification problems. For a good overview of the available techniques for protein structure prediction, we refer the readers to Kurgan and Homaeian (2006). The discriminating ability of the classifiers discussed above depends on different factors like data size, type and sample distribution. This is evident from the significant difference between the results for self-consistency and cross-validation tests performed by Chou (1995), Cai and Zhou (2000) and Cai et al. (2001). Some of these machine learning methods are known to be computationally intensive with the computational load depending on the number of classes or sample size. The superior methods like SVM are basically binary (two-class) classifiers and hence their extension to multi-class protein structure prediction involves formulation of many binary classifiers leading to immense computational efforts. Also, the existing methods do not capitalize on the association between the predictor features which have the potential to bring out distinct dissimilarities between classes. Such variable interactions characterizing the structure can be mathematically established and the distinct relations can be used as discriminating models. The new variable predictive model based class discrimination (VPMCD) algorithm proposed in this paper

[*] Corresponding author. Tel.: +65 6516 8484; fax: +65 6779 1936.
E-mail address: chels@nus.edu.sg (S. Lakshminarayanan).

1476-9271/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2008.03.009

attempts this new paradigm of model based data classification approach.

2. Methods

2.1. Problem Formulation: Protein Secondary Structure Prediction

In general, the problem can be defined as: can we predict the protein secondary structure given its primary sequence (characteristics polypeptide chain)? Every protein exhibits a unique secondary structure with distinct non-random features which are attributed to the orderly bonding of amino acid molecules. Hence, analyzing the amount and arrangement of each of amino acid molecules in the primary sequence forms the basis of any secondary/tertiary protein structure prediction algorithm. For statistical classification algorithms, many different continuous features are extracted from the set of given protein sequences and related to the respective structural classes. Most commonly, the molar composition of the 20 amino acids in the primary sequence is used as the basis feature vector for protein data analysis (Zhou, 1998). In general, any protein can be represented by a numerical vector of size p, where p is the number of features extracted from its alphabetical sequence. The dataset of n proteins and their corresponding amino acid compositions is denoted as matrix $N[n \times p]$. The main objective of any supervised structure prediction algorithm is to assign these proteins to different known structural classes (g). Hence the complete protein dataset for classification is denoted as $N[n \times p;\ g]$. To demonstrate the new classification algorithm in the present study, we select 20 amino acid compositions as feature set ($p = 20$) and classify the proteins into α, β, α/β and α + β groups ($g = 4$).

2.2. Concept of Variable Predictive Models

The characteristics of any multivariate system are mainly defined using a set of continuous variables and interactions between them. Some of the nonlinear associations and multivariate dependencies, common to biological systems, require richer quantitative representations and mathematical insights for characterizing certain definitive system behavior. Consider the data matrix N with amino acid compositions (columns of matrix N) as continuous variable vectors (X_i; $i = 1, 2, \ldots, 20$) describing the proteins, the system under study. Physically, the amino acid molecules combine with each other with definitive arrangement to provide one of the characteristic structures. These arrangements can be mathematically interpreted as the dependencies between different X_i and can be suitably modeled using the data in N for each of the four structures. Such models, representing the amino acid interactions are termed here as variable predictive models (VPMs). Any VPM$_i$ defined for a selected amino acid composition variable X_i is basically a parametric model (linear or nonlinear) developed with statistical rigor using available protein sequences. For a given set of proteins, the model VPM$_i$ can predict composition vector X_i using a best set of other compositions (X_j; $j \neq i$) extracted from the same set of proteins. The VPM$_i$ models attempt to mathematically mimic the inter-relationships between different amino acids which are presumably the basis for the different protein structures.

Though different forms of models can be designed and tested, simple polynomial models are adopted here to elucidate the VPM concept and its extension to a new supervised learning algorithm. The VPM$_i$ for a given amino acid composition variable (X_i) is obtained by selecting one of the two model types. These are polynomial (P), polynomial + interaction (PI) models as described by Eqs. (1) and (2). Linear and nonlinear effects are captured by deciding the order of the polynomial (l) in the models. The number of other amino acid compositions (size of variable set X_j; $j \neq i$) used for prediction in VPM$_i$ is referred as predictor variable number (r). Both univariate ($r = 1$) and multivariate ($r > 1$) models are used for the above two types of VPMs:

- Polynomial (P) VPM:

$$X_i = b_0 + \sum_{j=1}^{l}\sum_{k=1}^{r} b_{k\_j} X_k^j \quad \text{with } k \neq i \tag{1}$$

- Polynomial interaction (PI) VPM:

$$X_i = b_0 + \sum_{j=1}^{l}\sum_{k=1}^{r} b_{k\_j} X_k^j + \sum_{a=1}^{r}\sum_{b=a+1}^{r} b_{ab} X_a X_b \quad \text{with } k, a, b \neq i \tag{2}$$

The choice of model type (P or PI), polynomial order (l) and predictor variable number (r) is made and the parameters are estimated using the data available in the matrix N. One of the ways to determine the set of 'b' values is by formulating an ordinary least squares problem (Beck and Arnold, 1977) as $X_i = DB$, where B is the model coefficient vector (with size $q \times 1$) and D is the design matrix ($n \times q$) containing the polynomial values of predictor variable set as used in the right-hand side of Eqs. (1) and (2). The number of additive terms in the model and hence the number of coefficients (q) depends on the model type, l and r values selected during construction of VPM. For example, with $P(l = 1, r = 1)$ we will get linear (L) univariate VPM as $X_1 = b_0 + b_{2\_1}X_2$ having $q = 2$ parameters. Similarly with PI ($l = 2$, $r = 2$) we can generate quadratic interaction (QI) multivariate VPM written as $X_1 = b_0 + b_{2\_1}X_2 + b_{3\_1}X_3 + b_{2\_2}X_2^2 + b_{3\_3}X_3^2 + b_{23}X_2X_3$ with $q = 6$. In order to obtain statistically meaningful model parameters for the VPM, using $B = D^{-1}X_i$ (where D^{-1} is the inverse of design matrix D), we must have $n \geq q$, i.e. the number of samples must exceed the number of parameters to be estimated. This criterion can be a good starting point to decide the type and order of the models. The elements of VPM thus obtained can be stored as a structure containing model type, selected l and r, vector of B values and information about predicting variables X_j that make up the D matrix. The unique feature of the proposed modeling approach is to utilize each X variable as dependent variable and the remaining variables as predictor variables in order to establish the interactions in the system without any prior knowledge of associations. The focus in the present study is to adopt the concept of variable prediction models as an effective discriminating tool for data classification applications. To our best understanding, this quantitative model based approach is the first of its kind for protein structure prediction applications.

2.3. VPMCD as Applied to Protein Structure Classification

The underlying principle of VPM based class discrimination (VPMCD) algorithm is to build variable predictive models for all the features separately for each class and use their predictive ability to classify a new observation. The set of VPM$_i$ belonging to a given particular group k (denoted as VPM$_i^k$), uniquely characterize the variable associations for that group and hence have distinctly higher accuracies in predicting any sample belonging to group k. The sample prediction capabilities for each of VPM$_i^k$ are thus the primary discriminating criteria used in VPMCD. Fig. 1 gives the schematic work flow of the generalized VPMCD algorithm for protein secondary structure prediction. The training set N is sepa-

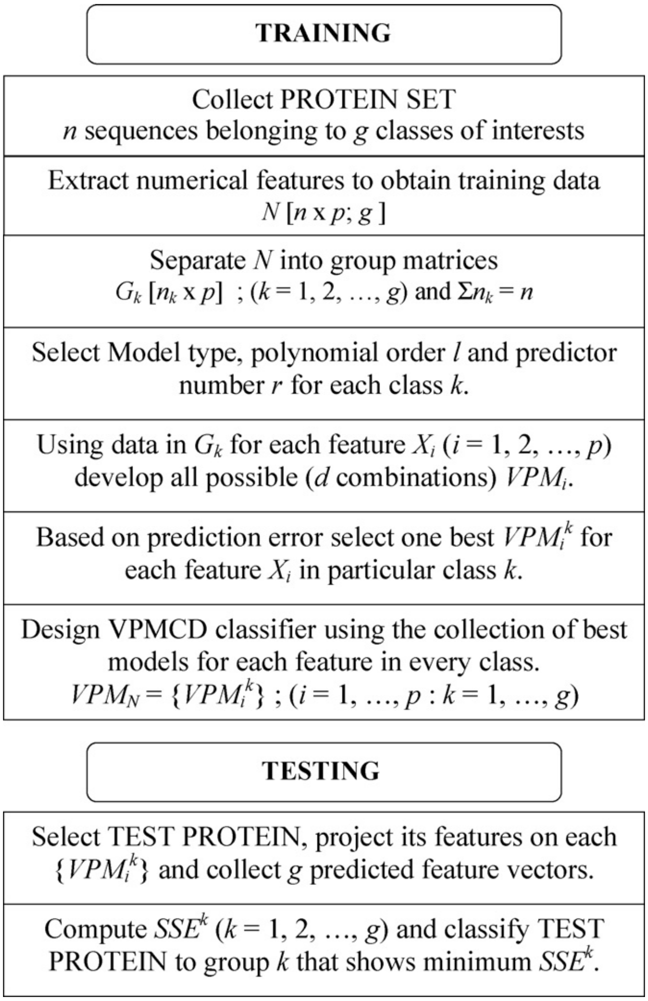

TRAINING
Collect PROTEIN SET
n sequences belonging to g classes of interests
Extract numerical features to obtain training data
$N\ [n \times p;\ g\ ]$
Separate N into group matrices
$G_k\ [n_k \times p]$; $(k = 1, 2, \ldots, g)$ and $\Sigma n_k = n$
Select Model type, polynomial order l and predictor number r for each class k.
Using data in G_k for each feature X_i $(i = 1, 2, \ldots, p)$ develop all possible (d combinations) VPM_i.
Based on prediction error select one best VPM_i^k for each feature X_i in particular class k.
Design VPMCD classifier using the collection of best models for each feature in every class.
$VPM_N = \{VPM_i^k\}$; $(i = 1, \ldots, p : k = 1, \ldots, g)$
TESTING
Select TEST PROTEIN, project its features on each $\{VPM_i^k\}$ and collect g predicted feature vectors.
Compute SSE^k $(k = 1, 2, \ldots, g)$ and classify TEST PROTEIN to group k that shows minimum SSE^k.

Fig. 1. Generalized VPMCD algorithm for protein classification.

rated into matrices for each class of proteins $G_k[n_k \times p]$ where n_k is the number of proteins in the training set belonging to class g $(\Sigma n_k = n)$. The model order l, predictor variable number r and the model type as explained in previous section are decided. For each amino acid feature vector X_i, d sets of remaining variables (feature vectors other than X_i) are selected [$d = (p-1)C_r = 19C_r$ in present study]. Depending on the model type and order (l) a design matrix D. For each protein structure group ($k = \alpha$, β, α/β or α + β) the amino acid associations are modeled separately by designing corresponding VPM (as explained in previous section) for each set of $X_i - X_j$ association. The sets of model coefficients B_i are evaluated for all the possible d models VPM_i for predicting feature vector X_i. The learnt VPM_i is then used to predict back X_i using all the d models to obtain $X_{i\_\mathrm{pred}.}$ The best model (out of d possible models) VPM_i^k is selected as final predictive model for X_i in group k, according to its prediction accuracy based on sum of squared prediction errors. At the end of this learning procedure (training step) for all the g groups, each amino acid composition X_i will have one best predictive model VPM_i and corresponding set of model parameters B_i. The models and the parameter sets for the entire system N with g groups are stored as class models set $\mathrm{VPM}_N = \{\mathrm{VPM}_i^k\}$. Since the VPMCD algorithm captures all the direct inter-variable relations, it is conjectured that the final VPM_N obtained on the training data represents protein structure specific amino acid interaction discriminatory model to be used for sample testing. The algorithm then statistically tests each new protein (S) by projecting the selected amino acid compositions of S on model D_i^k and predicting the amino acid feature vector $\hat{S}$ using corresponding VPM_i^k. The VPMCD classifier is then built using the objective function as given in Eq. (3). The protein S is classified as belonging to class k based on the minimum squared prediction error SSE^k:

$$\min_k ||\mathrm{SSE}^k|| = \sum_{j=1}^{p} (S_j - \hat{S}_j)^2; \quad k \in \{\alpha, \beta, \alpha/\beta, \alpha+\beta\} \tag{3}$$

The new VPMCD classification algorithm for protein structure prediction is built on the hypothesis that if the test protein belongs to one of the structures used during supervised training, then its amino acid compositions can be closely predicted by the VPM_i^k models belonging to that class. Since all the VPM_i^k are independently trained and they store the best relations between variables for every structure, the models themselves directly discriminate the group structures. As all the 20 amino acid compositions of the test protein are predicted and used for fitness evaluation, the VPM_i^k based structure prediction is robust and can distinguish specific group characteristics even if the proteins are closely located in the 20 amino acid descriptor space. Since the criteria for discrimination are the prediction errors obtained using VPM_i^k, the VPM based discrimination method does not suffer from the problem of inseparability of proteins based on distance measure. VPMCD algorithm scheme can directly attempt a multi-class problem without having to use combination of multiple binary classifiers or project variables on different vector space (unlike SVM). The new algorithm does not involve excess tuning of parameters, iterative, optimization algorithms (like ANN, SVM). We demonstrate these aspects using well-established protein structure datasets while benchmarking the VPMCD performance with existing methods.

3. Experiments

3.1. Datasets

Two widely studied protein datasets constructed by Zhou (1998) are used to demonstrate the performance of the proposed VPMCD algorithm. The proteins were extracted by Zhou from the SCOP data base (Murzin et al., 1995). The first dataset (SCOP277) contains 277 proteins [α = 70; β = 61; α/β = 81; α/β = 65] and the second dataset (SCOP498) consists of 498 proteins [α = 107; β = 126; α/β = 136; α/β = 129] including many of the proteins from SCOP277. These datasets have been analyzed before using various distance measures (Nakashima et al., 1986), component coupled based discrimination method (Chou, 1995; Zhou, 1998), neural networks (Cai and Zhou, 2000), support vector machines (Cai et al., 2001) and most recently with new method developed using rough sets (Cao et al., 2006). Hence these protein datasets provide a good case studies for benchmarking VPMCD algorithm.

The primary amino acid sequences for all the proteins in the two datasets are extracted from Protein Data Bank (Berman et al., 2000) using the online text-based search and retrieval system ENTREZ (http://www.ncbi.nlm.nih.gov/entrez/). For each of the (277 + 498) protein sequences in both datasets, 20 distinct features based on compositions for each of the amino acids are extracted using MATLAB code developed in-house. This procedure results in the training set $N \sim [277 \times 20;\ 4]$ for SCOP277 and $N \sim [498 \times 20;\ 4]$ for SCOP498 dataset. A separate dataset with 222 proteins (α = 37, β = 64, α/β = 57, and α + β = 64) present in SCOP498 with homology <40% with SCOP277 proteins are extracted to formulate a new test sample set $N_{\mathrm{sample}} \sim [222 \times 20;\ 4]$.

Table 1
Re-substitution (RS) test and Jackknife (JK) results for different algorithms

Dataset	Method	Test	α	β	α/β	α + β	Overall (%)
SCOP277	CCD	RS	95.7	93.4	95.1	92.3	94.2
		JK	84.3	82.0	81.5	67.7	79.1
	ANN	RS	98.6	93.4	96.3	84.6	93.5
		JK	68.6	85.2	86.4	56.9	74.7
	SVM	RS	100	100	100	100	100
		JK	74.3	82.0	87.7	72.3	79.4
	Rough sets	RS	100	100	100	100	100
		JK	77.1	77.0	93.8	66.2	79.4
	VPMCD[a]	RS	100	100	100	100	100
		JK	85.7	85.0	92.9	84.4	84.2
SCOP498	CCD	RS	95.8	95.2	94.9	95.4	95.8
		JK	93.5	88.9	90.4	84.5	89.2
	ANN	RS	100	98.4	96.3	84.5	94.6
		JK	86.0	96.0	88.2	86.0	89.2
	SVM	RS	100	100	100	100	100
		JK	88.8	95.2	96.3	91.5	93.2
	Rough sets	RS	100	100	100	100	100
		JK	87.9	91.3	97.1	86.0	90.8
	VPMCD[a]	RS	100	100	100	100	100
		JK	93.5	94.3	97.7	92.2	94.5

For each data set best RS and JK performances are underlined.
[a] VPM model type used is QI and predictor variable number $r = 4$.

In order to investigate the effect of homology on the performance of the new method, a low homology protein dataset (PDB25) with 1673 proteins having an average homology not more than 25% (Kurgan and Homaeian, 2006) is also studied separately.

3.2. VPMCD Implementation and Testing

The VPMCD algorithm discussed in Section 2 has been implemented and executed in MATLAB (2005). Options to decide the model type and order for each VPM_i to be used during training are available. Different, already available modules in MATLAB are used to determine the design matrix D_i. VPM model types for L, LI, QI and Q are tried with $r = 1–5$. The datasets are separately subjected to VPMCD training and the classifier performance is validated using re-substitution (RS), Jackknife (JK) (leave one out cross-validation—LOOCV), random sub-sampling (% random) and new sample (NS) tests. The performance is compared with existing methods using the best-reported literature result for individual methods.

4. Results and discussion

Individual (group-wise) and overall prediction results for the two SCOP datasets are presented in Table 1. The total classification results are indicated in the last column with the overall percentage of correct classifications for all the proteins sampled. Results for re-substitution test (shown as RS in Table 1) clearly indicate the complete performance of supervised learning algorithms. All the four individual structures can be fully recognized and predicted. For both the datasets, the new VPMCD method is fully self-consistent with the protein classes. The results are similar to the well-established SVM method and better than neural networks and CCD algorithms. This is inline with the strength of the new method to capture the amino acid interactions distinctly into $\{\mathrm{VPM}_i^k\}$ for predicting the characteristic groups. The LOOCV results (shown as JK in Table 1) provide better insights to the superiority of the proposed method. Compared to best available SVM and ANN methods, VPMCD algorithm efficiently predicts the untrained test samples during the Jackknife test. This indicates the stability of the model based approach for mixed homology protein datasets, compared to SVM and neural networks. The additional 5% of proteins correctly predicted for SCOP277 dataset compared to SVM method reveal the better efficiency of variable association model based training approach even with smaller training set. Consistent with any data driven approach VPMCD algorithm provides better results for α/β samples and SCOP498 dataset as they provide more samples during training step.

4.1. Sensitivity Analysis

Effect of different model structures is already accounted during design of VPMCD using different forms of VPMs (L, LI, Q and QI). A detailed analysis on the effect of predictor variable number (r) and sequence homology in sample dataset is also performed. Table 2 highlights the effect of r on the VPMCD performance using SCOP277 dataset. Most of the previous works on protein structures (Zhou, 1998; Cai et al., 2001; Cao et al., 2006) have studied only the self-consistency and LOOCV tests for all the methods. VPMCD algorithm attains higher accuracies for the re-substitution and LOOCV tests establishing its potential to learn and generalize protein structures as compared to existing methods. The accuracy is better than that of CCD and ANN methods with just a bivariate QI type VPM_i model. Higher order models provide further increments in the VPMCD performance finally matching the best performance of 100% for re-substitution and superior 84.2% for LOOCV test with $r = 4$. This improvement due to increase in number of other amino acids used to predict a feature reveals the multivariate interactions among the building blocks of peptide chains. The improvement in

Table 2
Effect of model order on VPMCD performance using QI type models on SCOP277 dataset

Order	RS	JK	10% random	20% random	NS
$r = 1$	84.13	68.27	70.37 ± 11.2	64.26 ± 7.7[a]	81.53
$r = 2$	94.83	79.71	73.70 ± 9.5	76.67 ± 5.4	84.23
$r = 3$	98.16	83.40	79.26 ± 7.4	79.82 ± 4.6	91.89
$r = 4$	100	84.2	84.07 ± 5.2	81.30 ± 4.0	92.34

[a] For the 10 and 20% random sampling methods, the reported results are average values obtained with 10 trials and are shown with ±standard deviations.

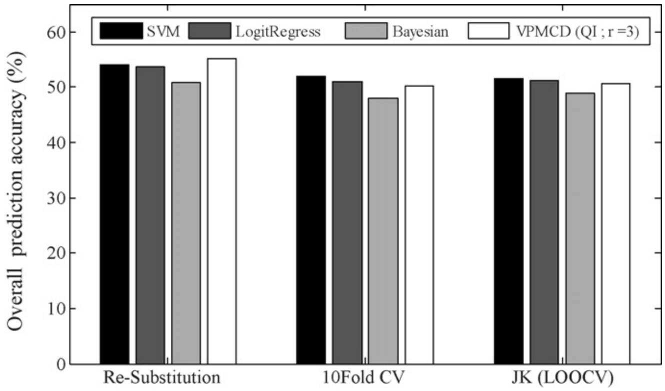


Fig. 2. VPMCD performance for low homology dataset (PDB25) compared with best results reported by Kurgan and Homaeian (2006).

prediction accuracies for other tests (multiple random sampling) is also evident with increase in order *r*. Decreasing standard deviation with higher *r* indicates the consistency and robustness of the VPMCD algorithm. The last column in Table 2 displays the performance of VPMCD in predicting a completely new sample set N_{sample}. 92.34% of the new proteins are correctly classified. Higher prediction using QI type models highlights the ability of VPMCD to capture the importance of higher order interactions between amino acids which was also observed by others (Chou, 1995; Edler et al., 2001). Fig. 2 outlines the results obtained for the low homology dataset (PDB25). Fixed set of models as used for high homology datasets (SCOP) are employed to build the classifier. In the present form the VPMCD provides similar performance compared to the reported best methods (Kurgan and Homaeian, 2006). These results bring out the fundamental advantage of the new method to capture the inherent protein structure in the form of variable association models. VPMCD algorithm is a potential tool for high homology datasets and with further tuning can be extended to low homology proteins. Effect of adding additional features and learning over larger domain of proteins, alternate forms of VPMs and their statistical validity can be further investigated. With these factors taken into consideration, the proposed new VPMCD algorithm comes out as a strong and potent tool for other data classification applications in computational biology.

References

Beck, J.V., Arnold, K.J., 1977. Parameter Estimation in Engineering and Science. Wiley, New York.

Berman, H.M., et al., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.

Cai, Y.D., Zhou, G.P., 2000. Prediction of protein structural classes by neural network. Biochimie 82 (8), 783–785.

Cai, Y.D., Liu, X.J., Xu, X.B., Zhou, G.P., 2001. Support vector machines for predicting protein structural class. BMC Bioinformatics 2, 3.

Cao, Y., Liu, S., Zhang, L., Qin, J., Wang, J., Tang, K., 2006. Prediction of protein structural class with rough sets. BMC Bioinformatics 7, 20.

Chou, K.C., 1995. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins Struct. Funct. Genet. 21, 319–344.

Edler, L., Grassmann, J., Suhai, S., 2001. Role and results of statistical methods in protein fold class prediction. Math. Comput. Model. 33, 1401–1417.

Kurgan, L.A., Homaeian, L., 2006. Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recogn. 39, 2323–2343.

MATLAB, 2005. MATLAB 7.0.4 Release 14. The MathWorks Inc., Natick, MA.

Murzin, A., Brenner, S., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of protein database for the investigation of sequence and structures. J. Mol. Biol. 247, 536–540.

Nakashima, H., Nishikawa, K., Ooi, T., 1986. The folding type of protein is relevant to the amino acid composition. J. Biochem. 99, 152–162.

Zhou, G.P., 1998. An intriguing controversy over protein structural class prediction. J. Protein Chem. 17, 729–738.