Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2008, Article ID 235451, 7 pages
doi:10.1155/2008/235451

Research Article

A Time-Series-Based Feature Extraction Approach for Prediction of Protein Structural Class

Ravi Gupta,[1,2] Ankush Mittal,[1] and Kuldip Singh[1]

[1] *Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee, Roorkee 247 667, Uttarakhand, India*
[2] *Information Science Division, AU-KBC Research Centre, Anna University, MIT Campus, Chennai 600 044, India*

Correspondence should be addressed to Ravi Gupta, ravigupta@au-kbc.org

Received 28 May 2007; Revised 21 November 2007; Accepted 10 March 2008

Recommended by Stéphane Robin

This paper presents a novel feature vector based on physicochemical property of amino acids for prediction protein structural classes. The proposed method is divided into three different stages. First, a discrete time series representation to protein sequences using physicochemical scale is provided. Later on, a wavelet-based time-series technique is proposed for extracting features from mapped amino acid sequence and a fixed length feature vector for classification is constructed. The proposed feature space summarizes the variance information of ten different biological properties of amino acids. Finally, an optimized support vector machine model is constructed for prediction of each protein structural class. The proposed approach is evaluated using leave-one-out cross-validation tests on two standard datasets. Comparison of our result with existing approaches shows that overall accuracy achieved by our approach is better than exiting methods.

1. Introduction

Determination of protein structure from its primary sequence is an active area of research in bioinformatics. The knowledge of protein structures plays an important role in understanding their functions. Understanding the rules relating the amino acid sequence to the three-dimensional structure of the protein is one of the major goals of contemporary molecular biology. However, despite more than three decades of both experimental and theoretical efforts prediction of protein structure still remains one of the most difficult issues.

The concept of protein structural classes was originally introduced by Levitt and Chothia [1] based on a visual inspection of polypeptide chain topologies in a dataset of 31 globular proteins. A protein (domain) is usually classified into one of the following four structural classes: all-α, all-β, α/β, and $\alpha + \beta$. Structural class categorizes various proteins into groups that share similarities in the local folding patterns. The all-α and all-β classes represent structures that consist of mainly α-helices and β-strands, respectively. The α/β and $\alpha + \beta$ classes contain both α-helices and β-sheets where the α/β class includes mainly parallel α-helices and β-strands and $\alpha + \beta$ class includes those in which α-helices and β-strands are largely segregated. Prediction of structural classes is based on identifying these folding patterns based on thousands of already categorized proteins, and applying these patterns to unknown structures but known amino acid sequences. Structural Classification of Proteins (SCOP) [2] is one of the most accurate classifications of protein structural classes and has been constructed by visual inspection and comparison of structures by experts.

In the past two decades several computational techniques for prediction of protein structural classes have been proposed. Prediction is usually a two-step process. In the first step a fixed length feature vector is formed from protein sequences which are of different length. The second step involves a classification algorithm. Klein and Delisi [3] proposed a method for predicting protein structural classes from amino acid sequence. Later on, Klein [4] presented a discriminant analysis based technique for this problem. Zhou et al. [5] in 1992 proposed a weighting method to predict protein structural class from amino acids. A maximum component coefficient method was proposed by

Zhang and Chou [6]. A neural network based approach [7] for protein structural classes was also developed using six hydrophobic amino acid patterns together with amino acid composition. A new algorithm that takes into account the coupling effect among different amino acid components of a protein by a covariance matrix is proposed in [8]. In [9], Chou and Zhang introduced Mahalanobis distance to reflect the coupling effect among different amino acids components, improving the accuracy of the current problem. A support vector machine (SVM) method using amino acid composition features for prediction of protein structural class was presented by Cai et al. [10] in 2001 and is one of the most accurate methods for classification. A supervised fuzzy clustering approach based on amino acid composition features was introduced by Shen et al. [11]. A combined approach, LogitBoost, was proposed by Feng et al. [12]. It combines many weak classifiers together to build a stronger classifier. In 2006, Cao et al. [13] proposed a rough set algorithm based on amino acid compositions and 8 physicochemical properties data.

In this paper, a three step procedure is proposed for prediction of protein structural class. The main contribution of this paper is in providing a novel feature vector which is obtained by applying a wavelet-based time-series analysis approach. The proposed feature extraction from protein sequence is inspired from the work of Vannucci and Lio [14] on transmembrane proteins. The fixed length feature vector for classification proposed is derived from ten physicochemical properties of protein sequences. The physicochemical properties are used to convert the protein sequences from symbolic domain to numeric domain and to derive a time series representation for protein sequences. Features are extracted by applying a wavelet-based analysis technique for time series data on mapped protein sequences. The feature vector summarizes the variation of physicochemical properties in the protein sequence. Finally, a support vector machine is trained using the novel feature vector and the parameters are optimized for generating accurate model (providing highest prediction accuracy).

Leave-one-out cross-validation also called jackknife test was performed on the datasets that were constructed by Zhou [15] from SCOP. The datasets were also used by Cai et al. [10], Cao et al. [13] for their experiments. An overall accuracy of 82.97% and 93.94% was achieved for 277 domains and 498 domains datasets, respectively, using the proposed approach.

The paper is organized as follows. In Section 2, we describe the steps followed for extracting wavelet variance features from protein sequences. A brief introduction to support vector machine (SVM) is also provided in this section. Section 3 provides the experiment results obtained for datasets of structural protein sequences. Conclusion follows in Section 4.

2. Method

The proposed approach for identification of structural classes of proteins is divided into three different stages: amino acid mapping, feature extraction, and classification. In the first stage the protein sequences are mapped to various physicochemical scales as provided in the literature. After this mapping procedure the protein sequences become discrete time series data. The second stage involves construction of fixed length feature vector for classification. The feature vector is generated by combining wavelet variance [16] features extracted from different physicochemical scales used for mapping stage. Finally, an SVM-based classification is performed based on the novel extracted features to identify the structural class of a protein sequence.

2.1. Amino Acid Mapping

In this stage, ten different physicochemical amino acid properties were used. The first is the average flexibility indices provided by Bhaskaran and Ponnuswamy [17]. The second is the normalized hydrophobicity scales provided by Cid et al. [18]. The third is the transfer free energy given by M. Charton and B. I. Charton [19] and cited by Simon [20]. The fourth is the residue accessible surface area in folded protein provided by Chothia [21]. The fifth is the relative mutability obtained by multiplying the number of observed mutations by the frequency of occurrence of the individual amino acids and is provided by Dayhoff et al. [22]. The sixth is the isoelectric point provided by Zimmerman et al. [23]. The seventh is the polarity of amino acids provided by Grantham [24]. The eight is the volume of amino acid provided by Fauchere et al. [25]. The ninth is the composition of the amino acids provided by Grantham [24]. The tenth is the molecular weight of the amino acids given by Fasman [26]. The numerical indices representing physicochemical property of amino acids were downloaded from http://www.genome.jp/dbget/.

2.2. Feature Construction

The representation of a protein sequence by a fixed length feature vector is one of the primary tasks of any protein classification technique. In this section, we present a wavelet-based time-series approach for constructing feature vector. Wavelet transform is a technique that decomposes a signal into several groups (vectors) of coefficients. Different coefficient vectors contain information about characteristics of the sequence at different scales. The proposed feature vector contains information about the variability of ten physiochemical properties of protein sequences over different scales. The variability of physiochemical properties is represented in terms of wavelet variance [16].

In the present work, a variation of the orthonormal dis-crete wavelet transform (DWT) [27, 28], called the maximal overlap DWT (MODWT) [29] is applied for feature extraction. In past, MODWT has been applied for analysis of atmospheric data [30] and economic time series data [31, 32]. The MODWT is a highly redundant and nonorthogonal transform. The MODWT was selected over DWT because it can handle any sample size N, while Jth order DWT restricts the sample size to multiple of 2^J. The property is very useful for analysis of protein sequences, as the length of the sequences is not a multiple of 2^J. In addition, MODWT yields an estimator of the variance of the wavelet coefficients that is statistically more efficient than the corresponding estimator based on the DWT.

Let $\mathbf{P}$ be an N-dimensional column vector containing the mapped protein sequence series $P_0, P_1, \ldots, P_{N-1}$, where N is the length of the protein sequence. It is assumed that P_t was collected at time $t\Delta t$, where Δt is the time interval between consecutive observation (in the present case Δt is equal to 1 amino acid). The MODWT of $\mathbf{P}$ for maximum level J is given by

$$\mathbf{Q} = \widetilde{\mathbf{W}}\,\mathbf{P}, \tag{1}$$

where $\mathbf{Q}$ is a column vector of length $(J+1)N$, and $\widetilde{\mathbf{W}}$ is an $(J+1)N \times N$ real-valued nonorthogonal matrix. The vector of MODWT coefficients given in (1) may be decomposed into $J+1$ vectors:

$$\mathbf{Q} = \left[Q_1, Q_2, \ldots, Q_J, R_J\right]^T, \tag{2}$$

where Q_j (where $j = 1, 2, \ldots, J$) and R_J are column vectors of length N. The vector Q_j contains the MODWT wavelet coefficients associated with change in $\mathbf{P}$ on scale of length $\tau_j = 2^{j-1}$, while R_J is a vector containing the MODWT scaling coefficients associated variation at scales of length 2^J and higher. In addition to MODWT coefficients, the matrix $\widetilde{\mathbf{W}}$ can be decomposed into $J+1$ submatrices, each of them $N \times N$ and is given by

$$\widetilde{\mathbf{W}} = \left[\widetilde{W}_1, \widetilde{W}_2, \ldots, \widetilde{W}_J, \widetilde{V}_J\right]^T. \tag{3}$$

Instead of using the wavelet and scaling filters, the MODWT utilizes the rescaled filters, that is, $\tilde{h}_j = h_j/2^j$ and $\tilde{g}_j = g_j/2^j$ (where, $j = 1, 2, \ldots, J$). The terms h_j and g_j are wavelet and scaling filters, respectively. The wavelet filter approximates high-pass filter, and the scaling filter approximates low pass filter. Details regarding wavelet and scaling filters can be found in [29]. The $N \times N$ dimensional submatrix $\widetilde{W}_1$ is constructed by circularly shifting the rescaled wavelet filter $\tilde{h}_1$ by integer units to the right so that

$$\widetilde{W}_1 = \left[\tilde{h}_1^{(1)}, \tilde{h}_1^{(2)}, \ldots, \tilde{h}_1^{(N-1)}, \tilde{h}_1^{(0)}\right]^T. \tag{4}$$

Similarly, $\widetilde{W}_2, \widetilde{W}_3, \ldots, \widetilde{W}_J$ can be obtained. The MODWT is an energy-preserving transform [29, 33] and is given as

$$\|\mathbf{P}\|^2 = \|\mathbf{Q}\|^2 = \sum_{j=1}^{J} \|Q_j\|^2 + \|R_J\|^2. \tag{5}$$

The sample variance (empirical power) of $\mathbf{P}$ is decomposed into pieces that are associated with scales $\tau_1, \tau_2, \ldots, \tau_J$

$$\hat{\sigma}_P^2 \equiv \frac{1}{N}\|\mathbf{P}\|^2 - \overline{P}^2 = \frac{1}{N}\sum_{j=1}^{J} \|Q_j\|^2 + \frac{1}{N}\|R_J\|^2 - \overline{P}^2, \tag{6}$$

Table 1: Dataset for the current study.

	all-α	all-β	α/β	$\alpha+\beta$	Total
Dataset1	69	61	81	65	276
Dataset2	105	126	135	129	495

where $\hat{\sigma}_P^2$ is the sample variance of $\mathbf{P}$, and $\overline{P}$ is its mean. The term $\|Q_j\|^2/N$ represents the contribution to the sample variance of $\mathbf{P}$ due to change at scale τ_j. For example, the average flexibility indices property of a protein sequence in terms of wavelet variance vector is given as follows:

$$\mathbf{F}_{\mathbf{Flex}} = \left[\hat{\sigma}^2_{\text{Flex}}(1), \hat{\sigma}^2_{\text{Flex}}(2), \ldots, \hat{\sigma}^2_{\text{Flex}}(J)\right]^T, \tag{7}$$

where J is the maximum level of decomposition of the time series data, that is, protein sequence. Similarly, wavelet variance vectors for hydrophobicity, transfer free energy, residue accessible surface area, relative mutability, isoelectric point, polarity, volume, composition, and molecular weight are calculated and are represented by $\mathbf{F}_{\mathbf{Hyp}}$, $\mathbf{F}_{\mathbf{Free}}$, $\mathbf{F}_{\mathbf{Area}}$, $\mathbf{F}_{\mathbf{Mut}}$, $\mathbf{F}_{\mathbf{Iso}}$, $\mathbf{F}_{\mathbf{Pol}}$, $\mathbf{F}_{\mathbf{Vol}}$, $\mathbf{F}_{\mathbf{Comp}}$, and $\mathbf{F}_{\mathbf{Mol}}$, respectively. The feature vector $\mathbf{F}_{\mathbf{PFold}}$ is constructed by concatenating all seven wavelet variance vectors and is given as follows:

$$\begin{aligned}\mathbf{F}_{\mathbf{PFold}} = {} & \mathbf{F}_{\mathbf{Flex}} \oplus \mathbf{F}_{\mathbf{Hyp}} \oplus \mathbf{F}_{\mathbf{Free}} \oplus \mathbf{F}_{\mathbf{Area}} \oplus \mathbf{F}_{\mathbf{Mut}} \oplus \mathbf{F}_{\mathbf{Iso}} \\ & \oplus \mathbf{F}_{\mathbf{Pol}} \oplus \mathbf{F}_{\mathbf{Comp}} \oplus \mathbf{F}_{\mathbf{Vol}} \oplus \mathbf{F}_{\mathbf{Mol}}.\end{aligned} \tag{8}$$

The physiochemical variation of a protein sequence is summarized in the proposed feature vector. The dimension of $\mathbf{F}_{\mathbf{PFold}}$ is equal to 10^*J and is dependent on the number of levels (J) to which the time series data (i.e., protein sequence) has to be decomposed. The value of J is further dependent on the length of time series data (i.e, protein sequence length) and $J \leq \log_2 N$, where N is the number of observation points in the time series or the length of protein. As most of the protein sequences taken up for the experiment have length greater than 32, we have selected $J = 5$. In this study, Daubechies [27] wavelet has been used for analysis.

2.3. Classification

The SVM was proposed by Cortes and Vapnik [34] as a very effective technique for pattern classification. SVM is based on the principle of structural risk minimization (SRM), which bounds the generalization error to the sum of training set error and a term depending on the Vapnik-Chervonenkis dimension [34] of the learning machine. The SVM induction principle minimizes an upper bound on the error rate of a learning machine on test data (i.e., generalization error), rather than minimizing the training error itself which is used in empirical risk minimization. This helps them to generalize well on the unseen data.

An open-source SVM implementation called LIBSVM [35] was used for classification. It provides various kernel types: radial basis function (RBF), linear, polynomial and sigmoid. Experiments were conducted using different kernels; however the RBF was selected because of its superior performance for the current work. Further, for finding the optimum values of parameters (C, γ) for RBF kernel, LIBSVM provides an automatic grid search technique using cross-validation. Basically various pairs of (C, γ) are tried

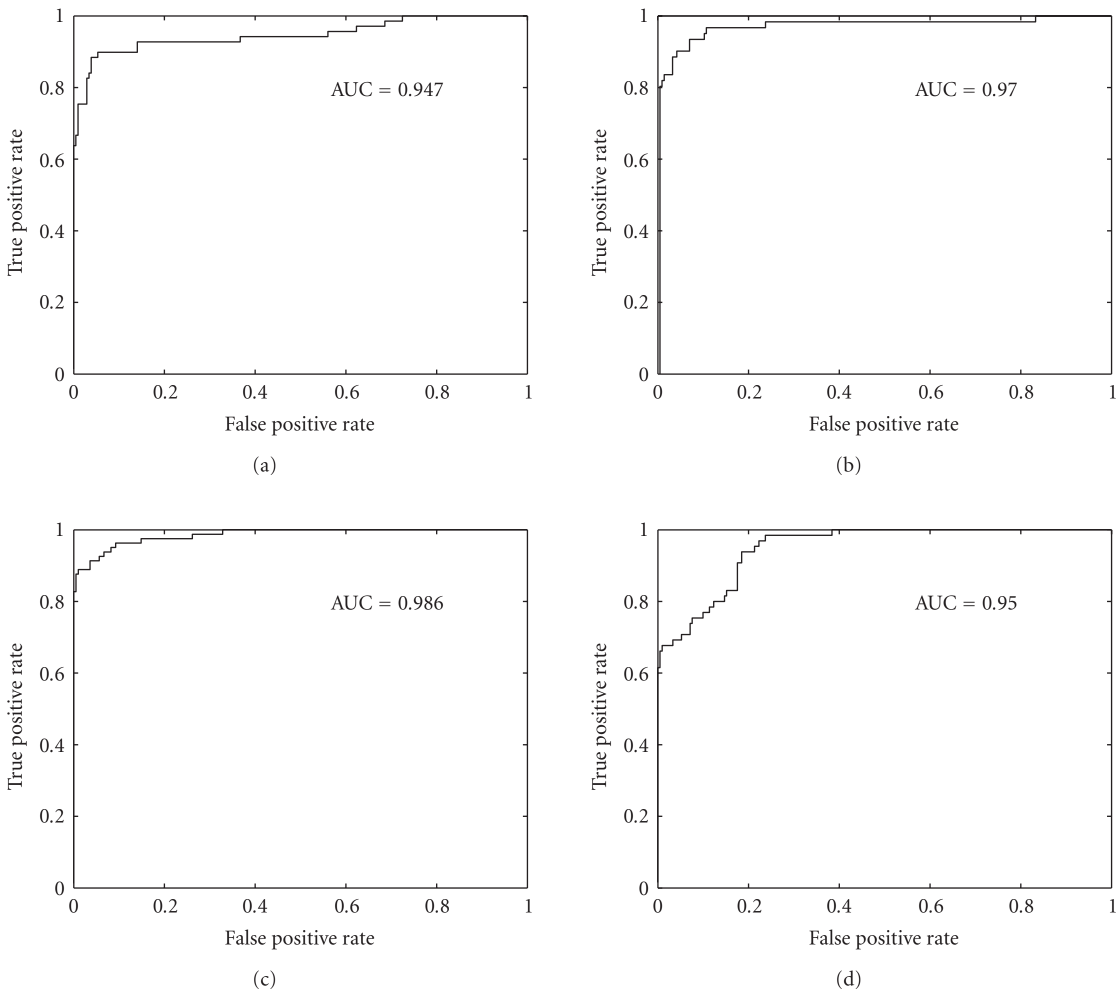

Figure 1: The ROC curve for identification of four structural classes of dataset1, all-α domains (a), all-β domains (b), α/β domains (c), and $\alpha + \beta$ domains (d).

Table 2: Experimental result of one-versus-others test on dataset1 evaluated using LOOCV.

	all-α	all-β	α/β	$\alpha + \beta$
True positive (TP)	60	54	72	43
False negative (FN)	9	7	9	22
True negative (TN)	199	208	191	209
False positive (FP)	8	7	4	2
(TP + TN)/(TP + FN + TN + FP) in %	93.84%	94.93%	95.29%	91.67%
Area under curve (AUC)	0.947	0.970	0.986	0
Optimal SVM parameters	$C = 20$ $\gamma = 0.07$	$C = 10$ $\gamma = 0.03$	$C = 10$ $\gamma = 0.5$	$C = 2$ $\gamma = 0.3$

and the one that provides best cross-validation accuracy is selected.

3. Experimental Results

To evaluate the performance of our approach two datasets of protein sequences constructed by Zhou [15] are used. The first dataset consists of 277 domains, of which 70 are all-α domains, 61 all-β domains, 81 are α/β domains, and 65 are $\alpha + \beta$ domains. The second dataset consists of 498 domains, of which 107 are all-α domains, 126 all-β domains, 136 are α/β domains, and 129 are $\alpha + \beta$ domains. The datasets were preprocessed before using for the experiment. The protein sequences having length less than 32 amino acids (as $J = 5$, where J is the maximum level of decomposition for wavelet transform) were removed from the dataset. The number of

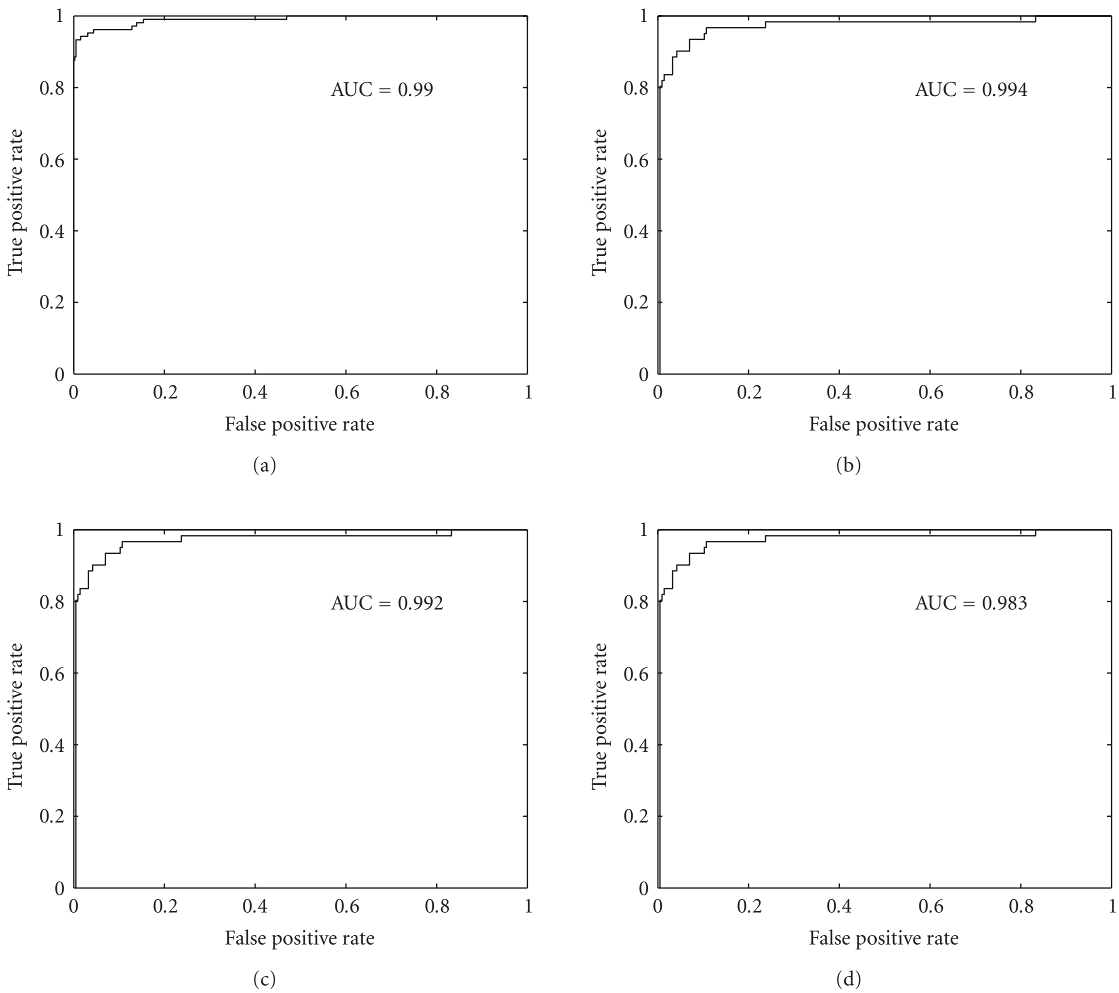

Figure 2: The ROC curve for identification of four structural classes of dataset2, all-α domains (a), all-β domains (b), α/β domains (c), and $\alpha + \beta$ domains (d).

protein sequences obtained after preprocessing both datasets is provided in Table 1.

The performance of the SVM classifier is measured using leave-one-out cross-validation (LOOCV) technique. LOOCV is n-fold cross-validation, where "n" is the number of instances in the datatset. Each instance in turn is left out, and the learning method is trained on all the remaining instances. It is judged by its correctness on the remaining instances-one or zero success or failure, respectively. The results of all "n" judgments, one for each member of the dataset, are averaged, and that average represents the final error estimate.

The classification of a protein sequence into one of the four structural classes is a multi-class classification problem. For identifying four different structural classes one-versus-others approach was followed. Four different SVMs were constructed, each specific to one class. The kth SVM was trained with all the samples of the kth class with positive labels and samples of remaining classes with negative labels. For example (Table 2, column 1 and Table 3, column 1), the SVMs for all-α domains protein sequences are positive labeled where as all-β domains, α/β domains, and $\alpha + \beta$ domains protein sequences are negative labeled. The experimental results obtained from the four SVMs for dataset1 and dataset2 are presented in Tables 2 and 3, respectively. The optimal SVM parameters obtained for the experiments are also provided in Tables 2 and 3. The accuracies for the current problem were calculated by applying the standard definition provided by previous work for multiclass protein sequence classification problem using SVM [36–38]. The prediction accuracy of the structural classes and overall prediction accuracy are given by

$$\text{Accuracy}_k = \frac{p_k}{\text{obs}_k}, \qquad \text{Overall accuracy} = \frac{\sum_{k=1}^{4} p_k}{M}, \tag{9}$$

where M is the total number of protein sequences, obs_k is the number of protein sequences of class "k," p_k is the number of correctly predicted protein sequences of class "k." The accuracy of each class and overall accuracy for

Table 3: Experimental result of one-versus-others test on dataset2 evaluated using LOOCV.

	all-α	all-β	α/β	$\alpha+\beta$
True positive (TP)	98	119	131	117
False negative (FN)	7	7	4	12
True negative (TN)	387	367	352	362
False positive (FP)	3	2	8	4
(TP + TN)/(TP + FN + TN + FP) in %	97.98%	98.18%	97.58%	96.77%
Area under curve (AUC)	0.990	0.994	0.992	0.983
Optimal SVM parameters	$C = 2$ $\gamma = 0.1$	$C = 2$ $\gamma = 0.1$	$C = 2$ $\gamma = 0.2$	$C = 1$ $\gamma = 0.2$

Table 4: Comparison of Leave-one-out cross-validation accuracy obtained for protein structural classification problem on the two datasets by our approach and existing approaches.

Dataset	Method	Prediction accuracy for each structural class (%)				Overall accuracy (%)
		all-α	all-β	α/β	$\alpha+\beta$	
Dataset1	Our approach	**86.96**	**88.52**	**88.89**	**66.15**	**82.97**
	Component coupled [6]	84.3	82.0	81.5	67.7	79.1
	Neural network [7]	68.6	85.2	86.4	56.9	74.7
	SVM [10]	74.3	82.0	87.7	72.3	79.4
	Rough sets [13]	77.1	77.0	93.8	66.2	79.4
Dataset2	Our approach	**93.33**	**94.44**	**97.04**	**90.7**	**93.94**
	Component coupled [6]	93.5	88.9	90.4	84.5	89.2
	Neural network [7]	86.0	96.0	88.2	86.0	89.2
	SVM [10]	88.8	95.2	96.3	91.5	93.2
	Rough sets [13]	87.9	91.3	97.1	86.0	90.8

datset1 and dataset2 calculated using (9) are shown in Table 4. The overall accuracy obtained by our approach for dataset1 and dataset2 is 82.97% and 93.94% respectively. The overall performance of our approach is better than existing techniques. Further, the receiver operating characteristic (ROC) curve and area under curve (AUC) for the proposed protein structural classification task were also calculated. An ROC curve is a plot of true positive rate as the ordinate versus the false positive rate as the abscissa; for a classifier, it is obtained by continuously varying the threshold associated with the decision function [39]. The ROC and AUC obtained for the one-versus-other experiment of dataset1 and dataset2 are presented in Figures 1 and 2, respectively. The ROC curve shown in Figure 1(a) is obtained when all-α domains protein sequences in dataset1 are positive labeled, where as all-β domains, α/β domains, and $\alpha+\beta$ domains protein sequences in dataset1 are negative labeled. Similarly, ROC curve for other classifications is also obtained.

4. Conclusion

In this work, we have presented a novel wavelet variance based feature vector for prediction of protein structural class. The aim of this research is to provide a new and complementary set of features for the current problem. Based on pattern recognition framework, the proposed approach is divided into three different tasks: amino acid mapping, feature construction, and classification. The feature vector summarizes the variation of ten different physicochemical properties of amino acids. The feature extraction technique is based on wavelet based time series analysis. Experiments were performed on two standard datasets (constructed by Zhou [15]). The result of LOOCV test shows that the proposed method achieves accuracy better than existing methods. The proposed approach can also be applied for identification of membrane protein type, enzyme family classification, and many others.

References

[1] M. Levitt and C. Chothia, “Structural patterns in globular proteins,” *Nature*, vol. 261, no. 5561, pp. 552–558, 1976.

[2] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, “SCOP: a structural classification of protein database for the investigation of sequence and structures,” *Journal of Molecular Biology*, vol. 225, no. 4, pp. 713–727, 1992.

[3] J. P. Klein and C. Delisi, “Prediction of protein structural class from the amino acid sequence,” *Biopolymers*, vol. 25, no. 9, pp. 1659–1672, 1986.

[4] P. Klein, “Prediction of protein structural class by discriminant analysis,” *Biochimica et Biophysica Acta*, vol. 874, no. 2, pp. 205–215, 1986.

[5] G. Zhou, X. Xu, and C.-T. Zhang, “A weighting method for predicting protein structural class from amino acid composition,” *European Journal of Biochemistry*, vol. 210, no. 3, pp. 747–749, 1992.

[6] C.-T. Zhang and K.-C. Chou, "An optimization approach to predicting protein structural class from amino acid composition," *Protein Science*, vol. 1, no. 3, pp. 401–408, 1992.

[7] B. A. Metfessel, P. N. Saurugger, D. P. Connelly, and S. S. Rich, "Cross-validation of protein structural class prediction using statistical clustering and neural networks," *Protein Science*, vol. 2, no. 7, pp. 1171–1182, 1993.

[8] K.-C. Chou, "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space," *Proteins: Structure, Function and Genetics*, vol. 21, no. 4, pp. 319–344, 1995.

[9] K.-C. Chou and C.-T. Zhang, "Predicting protein folding types by distance functions that make allowances for amino acid interactions," *Journal of Biological Chemistry*, vol. 269, no. 35, pp. 22014–22020, 1994.

[10] Y.-D. Cai, X.-J. Liu, X.-B. Xu, and G.-P. Zhou, "Support vector machines for predicting protein structural class," *BMC Bioinformatics*, vol. 2, article 3, pp. 1–5, 2001.

[11] H.-B. Shen, J. Yang, X.-J. Liu, and K.-C. Chou, "Using supervised fuzzy clustering to predict protein structural classes," *Biochemical and Biophysical Research Communications*, vol. 334, no. 2, pp. 577–581, 2005.

[12] K.-Y. Feng, Y.-D. Cai, and K.-C. Chou, "Boosting classifier for predicting protein domain structural class," *Biochemical and Biophysical Research Communications*, vol. 334, no. 1, pp. 213–217, 2005.

[13] Y. Cao, S. Liu, L. Zhang, J. Qin, J. Wang, and K. Tang, "Prediction of protein structural class with rough sets," *BMC Bioinformatics*, vol. 7, article 20, pp. 1–6, 2006.

[14] M. Vannucci and P. Lio, "Non-decimated wavelet analysis of biological sequences: applications to protein structure and genomics," *Sankhya B*, vol. 63, no. 2, pp. 218–233, 2001.

[15] G.-P. Zhou, "An intriguing controversy over protein structural class prediction," *Journal of Protein Chemistry*, vol. 17, no. 8, pp. 729–738, 1998.

[16] D. B. Percival, "On estimation of wavelet variance," *Biometrika*, vol. 82, no. 3, pp. 619–631, 1995.

[17] R. Bhaskaran and P. K. Ponnuswamy, "Positional flexibilities of amino acid residues in globular proteins," *International Journal of Peptide and Protein Research*, vol. 32, pp. 241–255, 1988.

[18] H. Cid, M. Bunster, M. Canales, and F. Gazitúa, "Hydrophobicity and structural classes in proteins," *Protein Engineering*, vol. 5, no. 5, pp. 373–375, 1992.

[19] M. Charton and B. I. Charton, "The structural dependence of amino acid hydrophobicity parameters," *Journal of Theoretical Biology*, vol. 99, no. 4, pp. 629–644, 1982.

[20] Z. Simon, *Quantum Biochemistry and Specific Interactions*, Abacus Press, Tunbridge Wells, Kent, UK, 1976.

[21] C. Chothia, "The nature of the accessible and buried surfaces in proteins," *Journal of Molecular Biology*, vol. 105, no. 1, pp. 1–12, 1976.

[22] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, "A model of evolutionary change in proteins," in *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, Ed., vol. 5, pp. 345–352, National Biomedical Research Foundation, Washington, DC, USA, 1978.

[23] J. M. Zimmerman, N. Eliezer, and R. Simha, "The characterization of amino acid sequences in proteins by statistical methods," *Journal of Theoretical Biology*, vol. 21, no. 2, pp. 170–201, 1968.

[24] R. Grantham, "Amino acid difference formula to help explain protein evolution," *Science*, vol. 185, no. 4154, pp. 862–864, 1974.

[25] J.-L. Fauchere, M. Charton, L. B. Kier, A. Verloop, and V. Pliska, "Amino acid side chain parameters for correlation studies in biology and pharmacology," *International Journal of Peptide and Protein Research*, vol. 32, no. 4, pp. 269–278, 1988.

[26] G. D. Fasman, *Practical Handbook of Biochemistry and Molecular Biology*, CRC Press, Boca Raton, Fla, USA, 1989.

[27] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, Pa, USA, 1992.

[28] S. G. Mallat, "Theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[29] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis*, Cambridge Press, Cambridge, UK, 2002.

[30] B. Whitcher, P. Guttorp, and D. B. Percival, "Wavelet analysis of covariance with application to atmospheric time series," *Journal of Geophysical Research*, vol. 105, no. D11, pp. 941–962, 2000.

[31] M. Gallegati and M. Gallegati, "Wavelet variance and correlation analyses of output in G7 countries," *Macroeconomics*, vol. 0512017, pp. 1–19, 2005.

[32] X. Xiong, X.-T. Zhang, W. Zhang, and C.-Y. Li, "Wavelet-based beta estimation of China stock market," in *Proceedings of the 4th International Conference on Machine Learning and Cybernetics (ICMLC '05)*, vol. 6, pp. 3501–3505, Guangzhou, China, August 2005.

[33] D. B. Percival and H. O. Mofjeld, "Analysis of subtidal coastal sea level fluctuations using wavelets," *Journal of the American Statistical Association*, vol. 92, no. 439, pp. 868–880, 1997.

[34] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[35] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," Tech. Rep., National Taiwan University, Taipei, Taiwan, 2004. http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[36] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[37] C. H. Q. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, no. 4, pp. 349–358, 2001.

[38] S. Hua and Z. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics*, vol. 17, no. 8, pp. 721–728, 2001.

[39] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.