Prediction of Protein Structural Class Using Novel Evolutionary Collocation-Based Sequence Representation

KE CHEN,[1] LUKASZ A. KURGAN,[1] JISHOU RUAN[2]
[1]*Department of Electrical and Computer Engineering, ECERF, University of Alberta, Edmonton, Alberta, Canada T6G 2V4*
[2]*Chern Institute of Mathematics, College of Mathematical Science and LPMC, Nankai University, Tianjin, People's Republic of China*

Received 24 May 2007; Revised 22 November 2007; Accepted 18 December 2007
DOI 10.1002/jcc.20918
Published online 21 February 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Knowledge of structural classes is useful in understanding of folding patterns in proteins. Although existing structural class prediction methods applied virtually all state-of-the-art classifiers, many of them use a relatively simple protein sequence representation that often includes amino acid (AA) composition. To this end, we propose a novel sequence representation that incorporates evolutionary information encoded using PSI-BLAST profile-based collocation of AA pairs. We used six benchmark datasets and five representative classifiers to quantify and compare the quality of the structural class prediction with the proposed representation. The best, classifier support vector machine achieved 61–96% accuracy on the six datasets. These predictions were comprehensively compared with a wide range of recently proposed methods for prediction of structural classes. Our comprehensive comparison shows superiority of the proposed representation, which results in error rate reductions that range between 14% and 26% when compared with predictions of the best-performing, previously published classifiers on the considered datasets. The study also shows that, for the benchmark dataset that includes sequences characterized by low identity (i.e., 25%, 30%, and 40%), the prediction accuracies are 20–35% lower than for the other three datasets that include sequences with a higher degree of similarity. In conclusion, the proposed representation is shown to substantially improve the accuracy of the structural class prediction. A web server that implements the presented prediction method is freely available at http://biomine.ece.ualberta.ca/Structural_Class/SCEC.html.

 J Comput Chem 29: 1596–1604, 2008

Key words: protein structure; domain structural class; PSI-BLAST; collocation of AA pairs; evolutionary information; SCOP, support vector machine

Introduction

Knowledge of protein structure plays a crucial role in protein function analysis, simulation of interaction between proteins and their ligands, rational drug discovery, and in many other applications. Numerous supplementary aspects of the protein structure, which include secondary structure, solvent accessibility, contact maps, fold, and structural class, are actively pursued in collaboration between bioinformaticians, structural biologists, and computer scientists. In this paper, we concentrate on the computational prediction of the structural classes.

The concept of protein structural class was proposed by Levitt and Chothia in 1976.[1] They inspected and classified 31 globular proteins into four structural classes: all-α, all-β, α/β, and $\alpha + \beta$. Nowadays, the most frequently used classifications of protein structural classes can be found in the structural classification of protein (SCOP) database.[2] This database is organized as a hierarchy of known protein and protein domain structures, in which the first level is based on the structural class. According to SCOP, the all-α and all-β classes correspond to structures that mainly consist of α-helices and β-strands, respectively. The proteins in the α/β and $\alpha + \beta$ classes contain both α-helices and β-strands; in the α/β class they are mainly interspersed, while in the $\alpha + \beta$ class they are segregated; see Figure 1.

Correspondence to: L. A. Kurgan; e-mail: lkurgan@ece.ualberta.ca

Contract/grant sponsor: Alberta Ingenuity Scholarship

Contract/grant sponsor: NSERC, Canada

Contract/grant sponsor: Liuhui Center for Applied Mathematics (China-Canada exchange program, MITACS)

Contract/grant sponsor: NSFC; contract/grant number: 100671100

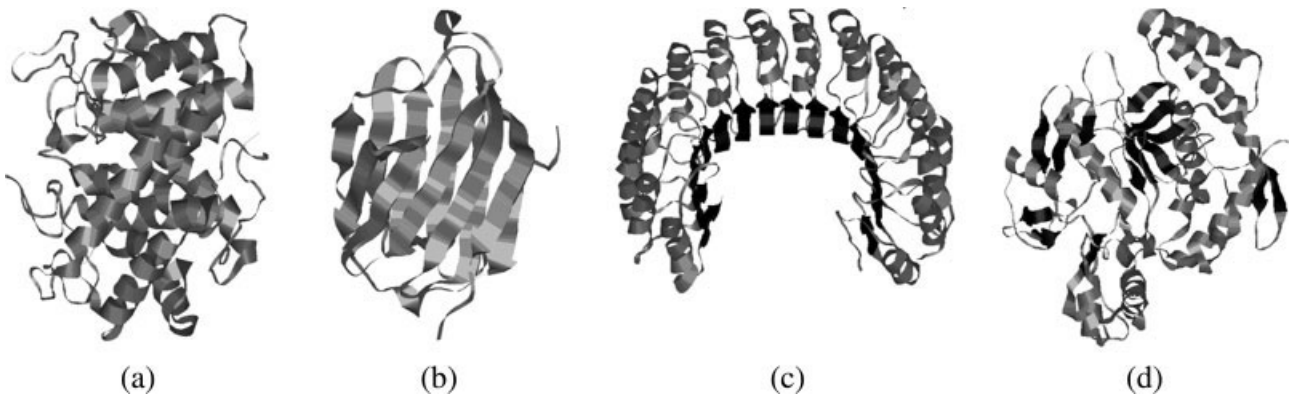

Figure 1. Ribbon drawings of representative protein structures that belong to the four structural classes: (a) all-α, (b) all-β, (c) α/β, (d) $\alpha + \beta$. In (c) and (d), strands are colored in black.

Knowledge of structural classes provides useful input for numerous applications that include prediction of protein unfolding rates,[3] prediction of DNA-binding sites,[4] discrimination of outer membrane proteins,[5,6] prediction of folding rates,[7] protein fold prediction,[8] secondary structure content prediction,[9,10] reduction of the conformation search space,[11] and for implementation of a heuristic approach to find tertiary structure.[12] At the same time, the structural class is known for relatively small number of proteins. The recent release 1.71 of SCOP database includes 75,930 protein domains, while release 22 of the NCBI's RefSeq database includes 3,438,099 known protein sequences. The main reason for such a wide gap is unavailability of protein structure, which is used to assign the corresponding structural class, for the significant majority of known protein sequences. Therefore, an accurate, automated method for classification of sequences into the corresponding structural classes would provide needed help in the laborious task of populating the SCOP database.

The last 20 years observed significant efforts in computational (automated) prediction of protein structural classes. Several early attempts were made in mid-1980s.[13,14] Composition vectors, autocorrelation function based on nonbonded residue energy, polypeptide composition, pseudo amino acid (AA) composition, and complexity measure factor were applied to represent protein sequence in later works.[15–20] Different classification algorithms, including the maximum component coefficient,[21] least correlation angle,[22] fuzzy clustering,[23] neural network,[24] Bayesian classification,[25] rough sets,[26] and component-coupled[15] and support vector machine (SVM),[27] have been already used to implement the structural class prediction methods. Recent works also explored application of complex classification models, such as ensembles,[28] bagging,[29] and boosting.[18,30] More details can be found in a recent review by Chou.[31]

While many recent works concentrate on applications of various, state-of-the-art classification algorithms, we concentrate on development of a novel representation of protein sequences. Although virtually all suitable classification algorithms have been already tried, design of suitable sequence representations received less attention. At the same time, researchers have shown that better-designed sequence representations lead to higher prediction accuracy,[31–33] which motivates our work. To this end, we propose a novel representation that is based on a PSI-BLAST profile,[34] which in turn incorporates evolutionary information with respect to the predicted sequence. The PSI-BLAST profile has been successfully applied in window-based protein prediction tasks, which include secondary structure, solvent accessibility, and transmembrane protein topology predictions,[35-37] while it was never applied to predict the structural class. This paper proposes a novel method that transforms the original profile, i.e., $N \times 20$ matrix, where N is the sequence length into a fixed length feature-vector based on a recently proposed collocation of AA pairs[38,39] that is calculated directly from the sequence. This proposed, novel representation is shown to substantially improve the accuracy of the structural class prediction.

Materials and Methods

Datasets

In contrast to majority of published methods that use one or two test datasets, we selected five widely used benchmark datasets to provide a comprehensive and unbiased comparison with previous studies. Two datasets were originally generated by Zhou[15] and were used in several past studies.[15,24,26,27,30] They include 277 and 498 protein domains, respectively. Another dataset, which consists of 204 domains, was generated by Chou[40] and further studied in a few publications.[18,20,23,40–45] The largest dataset, which was generated and studied in refs. 25, 28, and 32, includes 1189 domains. This dataset was generated from PDB40D_1.37 database in SCOP, and the sequence identity between any two sequences in this dataset is below 40%. The fifth dataset, which was created by Wang and Yuan,[25] consists of 675 domains, where each of these domains belongs to a different SCOP family. The sequence identity between any pair of sequences in this dataset is below 30%. Using current version of SCOP, 654 domains were found and annotated among the original 675 domains, and these domains constitute the fifth test set. Additionally, we selected a subset of these 654 domains that is

characterized by a lower maximal sequence identity that is set at 25% to create the sixth test set. We applied Smith–Waterman algorithm[46] to compute pairwise alignment for each pair of sequences in the dataset, and removed sequences that share more than 25% sequence identity. As a result, the final dataset includes 640 domains that share sequence identity below 25%. The dataset is available at http://biomine.ece.ualberta.ca/Structural_Class/SCEC.html. This set allows testing the quality of the proposed sequence representation for sequences characterized by very low similarity. All six datasets are balanced, i.e., each class includes similar number of sequences, and the corresponding class labels were extracted from SCOP.

Proposed Sequence Representation

The new representation, which combines PSI-BLAST profile and the concept of frequency of *collocation of AA pairs* in the sequence,[38,39] was developed for the proposed prediction method.

The original motivation to introduce the collocation of AA pairs comes from an insufficient sequence representation that is offered by the commonly used AA *composition vector*, which merely counts the frequencies of individual AAs in the sequence. At the same time, frequencies of AA pairs (dipeptides) provide more information, since they may reflect local (with respect to the sequence) interaction between AA pairs. Based on this argument, we should count all dipeptides in the sequence. Since there are 400 possible AA pairs (AA, AC, AD, ... , YY), a feature vector of that size is used to represent occurrence of these pairs in the sequence. At the same time, prior results show that short-range interactions between AAs, rather than interactions only between immediately adjacent AAs, have an impact of folding.[47] As a result, the proposed representation also considers collocated pairs of AAs, i.e. pairs that are separated by p other AAs. These pairs can be understood as the dipeptides with gaps. Collocated pairs for $p = 0, 1, \ldots, 4$ are considered, where for $p = 0$ the pairs reduce to the dipeptides. There are 400 feature values for each value of p.

On the other hand, numerous successful applications of PSI-BLAST profile illustrate that the evolutionary information is more informative than the sequence itself.[35–37] PSI-BLAST aligns a given query sequence to a database of sequences, and searches for these that are similar to the query sequence. Using multiple alignment, PSI-BLAST generates the frequency of each AA at each position in the query sequence. The PSI-BLAST profile generates 20-dimensional vector of AA frequencies for each position in the query sequence, which can be used to identify the key positions of conserved AAs and the residues that undergo mutations.

Our approach combines the frequency of collocation of AA pairs and the PSI-BLAST profile into so-called PSI-BLAST profile-based collocation of AA pairs. The PSI-BLAST profile is the $N \times 20$ matrix, which is denoted as $[a_{i,j}]$, where $i = 1, 2, \ldots, N$ denotes the position in the query sequence and $j = 1, 2, \ldots, 20$ denotes a given AA. After applying the substitution matrix and log function, a_{ij} values range between -9 and 11. The proposed representation is related to calculation of the frequency of AA pairs based on binary coding. The binary coding uses a 20-dimensional vector to encode each AA. The 20 AAs can be represented as AA_1, AA_2, ... , AA_{19}, and AA_{20}. In binary coding, AA_i is encoded as (0, 0, ... , 0, 1, 0, ... , 0, 0), where only the ith value is greater than 0. The binary coding matrix is denoted as $[b_{i,j}]$. The binary encoding and PSI-BLAST profile matrices have the same dimensionality ($N \times 20$).

The frequency of AA pairs can be computed from the binary coding matrix. For a given protein sequence $A_1A_2 \ldots A_N$.

A_iA_{i+1} is a AA_mAA_n dipeptide

$\iff A_i = AA_m$ and $A_{i+1} = AA_n$

$\iff b_{i,m} = 1$, $b_{i+1,n} = 1$, $b_{i,p} = 0$, $b_{i+1,q} = 0$, where $p \neq m$ and $q \neq n$

Given that $c_{s,t} = \min(b_{i,s}, b_{i+1,t})$, then

$$c_{s,t} = \begin{cases} 1 \ (\text{if} s = m, t = n) \\ 0 \ (\text{else}) \end{cases}$$

which means that AA_mAA_n was counted once while other dipeptides were counted 0 times. Matrix $[c_{s,t}]$ stores the frequencies of all dipeptides. The count of the AA pairs along the entire sequence can be computed as

$$c_{s,t} = \sum_{i=1}^{N-1} \min(b_{i,s}, b_{i+1,t})$$

The PSI-BLAST profile-based collocation of AA pairs is calculated in a similar way. The only difference is that the binary coding matrix $[b_{i,j}]$ is replaced by the PSI-BLAST profile $[a_{i,j}]$. The frequency of dipeptide AA_sAA_t is computed as $c_{s,t} = \sum_{i=1}^{N-1} \min(a_{i,s}, a_{i+1,t})$ and matrix $[c_{s,t}]$ stores the frequencies of all dipeptides.

Since the PSI-BLAST profile values can be negative while the frequencies of AA pairs should not be negative, the use of $\min(a_{i,s}, a_{i+1,t})$ function to represent the frequency of AA pairs is unsound. Instead, we use

$$c_{s,t} = \sum_{i=1}^{N-1} \max\left(0, \min(a_{i,s}, a_{i+1,t})\right)$$

in which the negative value of $\min(a_{i,s}, a_{i+1,t})$ is replaced by 0. Similarly, the frequencies of p-collocated AA pairs are defined as

$$d_{s,t,p} = \sum_{i=1}^{N-p-1} \max\left(0, \min(a_{i,s}, a_{i+p+1,t})\right)$$

The matrixes $[c_{s,t}]$ and $[d_{s,t,p}]$, which correspond to the frequency of the PSI-BLAST profile-based dipeptides and p-collocated AA pairs, respectively, constitute the proposed protein sequence representation. We generate PSI-BLAST profile-based collocation of AA pairs for $p = 0$, 1, 2, 3, and 4, which results in 2000 features for each sequence.

Table 1. List of Features Selected from the Set of PSI-BLAST Profile-Based Collocation of AA Pairs for $p = 0$, 1, 2, 3, and 4.

	PSI-BLAST profile-based collocation of AA pairs													
$p = 0$	AI	AL	AM	IA	II	IV	LA	LI	LV	MA	VI	VV		
$p = 1$	AE	RL	DI	DL	EI	EL	EM	II	IV	LE	MR	VI	VL	VV
$p = 2$	AI	AL	AM	GV	IA	II	IL	IM	IV	LA	LL	MA	VI	VV
$p = 3$	AA	IA	LL											
$p = 4$	AI	AL	AM	IA	LA	MA	VA							

The feature selection was performed using the dataset with 1189 domains.

Feature Selection

Since the proposed representation includes relatively large number of features, a feature selection method was used to reduce the dimensionality and potentially improve the prediction accuracy. Similar to ref. 39, the entropy-based feature selection method was used. This method evaluates each feature by measuring the information gain (IG) with respect to the class. The entropy of a feature X is defined as

$$H(X) = -\sum_i P(x_i)\ \log_2(P(x_i))$$

where $\{x_i\}$ is a set of values of X and $P(x_i)$ is the prior probability of x_i. The conditional entropy of X, given another feature Y (in our case the structural class) is defined as

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j)\ \log_2\big(P(x_i|y_j)\big)$$

where $P(x_i|y_j)$ is the posterior probability of X given the value y_i of Y. The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called IG.

$$\mathrm{IG}(X|Y) = H(X) - H(X|Y)$$

According to this measure, Y has a stronger correlation with X than with Z if $\mathrm{IG}(X|Y) > \mathrm{IG}(Z|Y)$. The feature selection was performed using 10-fold cross validation to avoid overfitting. In each fold, features were ranked based on their IG values (large IG value corresponds to a lower rank), and the final rank is computed as the average over the 10 folds. The best 50 features with the lowest average ranks were selected.

Since datasets with 204, 277, and 498 domains are characterized by high sequence identity and even share some duplicate sequences, the top 50 features for these three datasets are different when compared with the top features found for the remaining datasets. The three remaining datasets are characterized by lower, controlled level of sequences identity (i.e., 40%, 30%, and 25%), and as a result, the same group of features was applied in their classification. These 50 features were selected using the dataset with 1189 domains and are listed in Table 1.

Results and Discussion

Experimental Setup

The classification algorithms used to compare the proposed representation were implemented in Weka.[48] The representation, which includes 50 features, was tested with several representative state-of-the-art classifiers such as SVM,[49] multiple logistic regression,[50] nearest neighbor-based algorithm (IB1),[51] naïve Bayes,[52] and C4.5 decision tree.[53] These classification results, which were obtained for six benchmark datasets, were also compared with previous studies that used the same datasets and different sequence representations and classification algorithms. All experiments were performed using jackknife test and report the overall classification accuracy, as well as the accuracy for each of the four structural classes.

Experimental Results

Table 2 shows the classification accuracies when using the proposed sequence representation and the five selected classification algorithms. Among the five classifiers, SVM ranks the best for five datasets (204, 277, 1189, 675, and 640 domains) and IB1 ranks the best for two datasets (277 and 498 domains). The average overall accuracies of SVM for the six datasets are 78.0%, which is 4–10% higher than the average accuracies of the other four classifiers. The accuracies for the datasets with 204, 277, and 498 domains are much better than the results for the other three datasets. The reason for this difference is that the sequences in the datasets with 1189, 675, and 640 domains have lower sequence identity, i.e., below 40%, 30%, and 25%, respectively, while the other datasets include similar/duplicate sequences. The nearest neighbour classifier (IB1) performs well only for datasets with higher identity, since it relies on similarity between test and training sequences. In contrast, SVM is shown to perform well, irrespective of the underlying sequence similarity.

Among the four structural classes, $\alpha + \beta$ is the most difficult to predict. The average accuracy for the $\alpha + \beta$ class over the six datasets is 55.7%, which is 18–26% lower than the accuracies for the other three structural classes. Following, we demonstrate that a potential reason for such difference is associated with the relatively large variability of helix and strand content for proteins in the $\alpha + \beta$ class when compared with the other classes. At the same time, the proposed feature representation is shown to be correlated with the content of the secondary structures. Our classification results show that domains that belong to

Table 2. Comparison of Jackknife Accuracies Between Different Classification Algorithms That Use the Proposed Sequence Representation.

Dataset	Algorithm	Jackknife accuracy (%)				
		All-α	All-β	α/β	α + β	Overall
204 domains	SVM	90.38	100	91.11	93.48	**94.12**
	IB1	84.62	100	91.11	80.43	89.71
	C4.5	78.85	96.72	82.22	76.09	84.31
	Naïve Bayes	80.77	96.72	88.89	82.61	88.23
	Logistic regression	88.46	96.72	84.44	71.74	86.27
277 domains	SVM	91.18	91.38	93.42	76.92	**87.73**
	IB1	89.71	88.14	92.21	80.00	**87.73**
	C4.5	73.53	74.58	79.22	73.85	75.46
	Naïve Bayes	67.65	77.97	85.71	66.15	74.72
	Logistic regression	76.47	79.66	87.01	64.62	77.32
498 domains	SVM	97.98	93.33	95.62	93.43	94.93
	IB1	94.95	95.83	97.81	94.16	**95.74**
	C4.5	89.90	89.17	94.89	91.24	91.48
	Naïve Bayes	80.81	92.50	94.89	82.48	88.03
	Logistic regression	95.96	95.83	94.16	90.51	93.91
1189 domains (40% sequence identity)	SVM	75.80	75.18	82.57	31.78	**67.63**
	IB1	65.30	67.73	79.93	40.68	64.65
	C4.5	55.71	56.38	65.13	30.51	52.93
	Naïve Bayes	47.03	71.99	82.12	16.95	57.06
	Logistic regression	73.06	76.24	69.74	28.81	62.92
675 domains (30% sequence identity)	SVM	74.31	59.62	79.66	34.46	**61.47**
	IB1	54.86	47.44	68.93	35.03	51.53
	C4.5	54.86	50.64	57.06	33.33	48.62
	Naïve Bayes	55.56	62.18	81.92	21.47	55.05
	Logistic regression	69.44	62.18	60.45	33.33	55.50
640 domains (25% sequence identity)	SVM	73.91	61.04	81.92	33.92	**62.34**
	IB1	53.62	46.10	68.93	34.50	50.94
	C4.5	59.42	49.35	58.19	28.65	48.44
	Naïve Bayes	55.07	62.34	80.26	19.88	54.38
	Logistic regression	69.57	58.44	61.58	29.82	54.06

The best results are shown in bold.

α + β class were misclassified as the α/β class, as well as the all-α class and all-β classes. In dataset with 1189 domains, 14.8%, 28.8%, and 24.6% of the domains that belong to α + β class were misclassified as all-α, all-β, and α/β classes, respectively. In the dataset with 675 domains, the corresponding misclassification rates are similar and equal to 18.6%, 23.2%, and 23.7%, respectively. This shows that the lower quality on classification for α + β class is not only a result of an overlap between this class and the α/β class. In fact, the distribution of the secondary structure content, i.e., amount of helices and strands in the sequence, of all-α class, all-β class, and α/β class forms compact clusters that are relatively well separated from each other; see Figure 2a. In contrast, for the α/β class the distribution of the secondary structure content is relatively sparse and shares a more substantial overlap with the remaining classes; see Figure 2b. The figures indicate that the actual secondary structure content is not sufficient to distinguish α + β class among the other three classes and especially from the all-β and α/β classes. On the other hand, the features proposed in this work are correlated with the secondary structure content, see Table 3. We show that among the 50 features, 13 have the correlation coefficient value above 0.26 with the helix content, while 11 other features have the correlation coefficient value greater than 0.09 with strand content. Most importantly, the features that are positively correlated with the helix content are negatively correlated with the strand content and vice versa, which suggests that the computed features are associated with the underlying secondary structure content values. Finally, the 13 pairs that are positively correlated with helix content include only six amino acids, i.e., E, L, M, A, D, and R, and five of them (A, E, L, M, R) are shown to have the probability of above 0.5 to form helical structures.[47] The 11 pairs that are positively correlated with the strand content mainly include V and I amino acids, i.e., V occurs 12 times and I occurs 8 times in these pairs while the all other amino acids occur only twice. At the same time, these two residues are characterized by the largest probabilities to form strands.[47] Since the features generated in the paper reflect the secondary structure content and the secondary structure content is not sufficient to accurately distinguish α + β among the other three classes, the proposed features result in the lowest classification accuracy for this class.

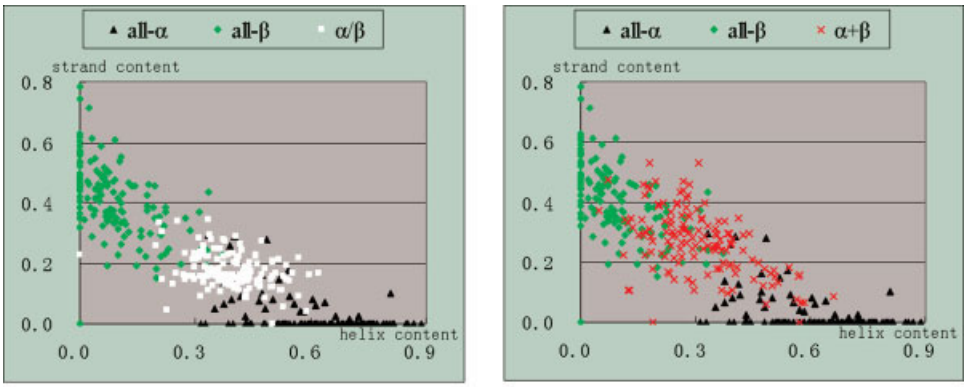

Figure 2. The distribution of the secondary structure content for the four structural classes in the dataset with 640 domains; *x*- and *y*-axis show helix and strand content, respectively. (a) The secondary structure contents of all-α, all-β, and α/β classes; the figure shows compact clusters for each structural class that are also well separated from each other. (b) The distribution of the secondary structure content of all-α, all-β, and $\alpha + \beta$ classes; the figure shows a sparse cluster for the $\alpha + \beta$ class, which also overlaps with the clusters for the all-α and all-β classes.

The proposed sequence representation and the best-performing SVM and IB1 classifiers were further compared with other recently reported prediction methods. The comparison includes recently reported results, which apply various classification algorithms and sequence representation for the five benchmark datasets, see Table 4. For the datasets with 277 and 498 domains, our results were compared with existing methods based on rough sets, component-coupling, neural network, SVM, and ensemble of boosted logistic regression classifiers.[15,24,26,27,30] For the dataset with 204 domains, the comparison group includes augmented covariant discriminate algorithm, unsupervised fuzzy clustering, supervised fuzzy clustering, SVM, increment and diversity algorithm, and ensemble of boosted logistic regression classifiers.[18,20,23,40–45,52] For the dataset with 1189 domains, our results are contrasted with Bayesian classifier, logistic regression and SVM, and ensemble of logistic regression, SVM, instance-based, and random forest classifiers.[25,28,32] Finally, our results are compared with a Bayesian classifier for the dataset with 675 domains.

For the datasets with 277 and 498 domains, the usage of the proposed representation results in substantial error rate reduction, i.e., 3.6/15.9 = 23% and 0.9/5.2 = 17%, respectively, when compared with the best previously reported results that were obtained by the LogitBoost classifier that uses AA compo-

Table 3. List of Proposed Features That Are Characterized by the Strongest Positive Correlation with Helix and Strand Content.

Features with highest correlation with helix content			Features with highest correlation with strand content		
Pair	Correlation coefficient with helix content	Correlation coefficient with strand content	Pair	Correlation coefficient with helix content	Correlation coefficient with strand content
E*L	0.358	−0.335	V*V	−0.221	0.286
L*E	0.351	−0.311	V*I	−0.205	0.265
L***L	0.331	−0.299	G**V	−0.245	0.259
E*M	0.324	−0.296	I*V	−0.187	0.254
L**L	0.295	−0.268	I*I	−0.163	0.222
AM	0.279	−0.259	V*L	−0.137	0.221
AL	0.275	−0.257	VV	−0.153	0.189
D*L	0.273	−0.287	IV	−0.118	0.151
L**A	0.273	−0.265	VI	−0.120	0.151
R*L	0.267	−0.246	V**V	−0.052	0.092
A****M	0.265	−0.262	II	−0.064	0.090
A****L	0.263	−0.255			
M**A	0.261	−0.258			

*Represents a gap between the corresponding pair of AAs.

Table 4. Comparison of Jackknife Accuracies Between the Two Best Classifiers That Use the Proposed Representation and Other Reported Methods.

Dataset	Algorithm	Reference	Feature-based sequence representation	Jackknife accuracy (%)				
				All-α	All-β	α/β	α + β	Overall
204 domains	Augmented covariant discriminate algorithm	20	Pseudo AA composition and complexity measure factor	82.7	90.2	100	87.0	89.7
	Unsupervised fuzzy clustering	41	AA composition	67.3	86.9	46.7	60.9	68.1
	Supervised fuzzy clustering	23	AA composition	73.1	90.2	62.2	63.1	73.5
	LogitBoost	18	AA composition	90.4	88.5	80.0	73.9	83.8
	BTSVM	45	Pseudo AA composition	90.4	100	73.9	97.8	91.2
	SVM	42	Pair-coupled amino acid composition	75	90	64	64	74.5
	SVM	44	Pseudo AA composition	88.5	96.7	77.8	73.9	85.3
	IDQD	43	AA composition and dipeptides	90.4	93.4	100	89.1	93.1
	SVM	This paper	PSI-BLAST based p-collocated AA pairs	90.4	100	91.1	93.5	**94.1**
	IB1	This paper	PSI-BLAST based p-collocated AA pairs	84.6	100	91.1	80.4	89.7
277 domains	Rough sets	26	AA composition and physicochemical properties	77.1	77.0	93.8	66.2	79.4
	Component-coupling	15	AA composition	84.3	82.0	81.5	67.7	79.1
	Neural network	24	AA composition	68.6	85.2	86.4	56.9	74.7
	SVM	27	AA composition	74.3	82.0	87.7	72.3	79.4
	LogitBoost	30	AA composition	81.4	88.5	92.6	72.3	84.1
	SVM	This paper	PSI-BLAST based p-collocated AA pairs	91.2	91.4	93.4	76.9	**87.7**
	IB1	This paper	PSI-BLAST based p-collocated AA pairs	89.7	88.1	92.2	80.0	**87.7**
498 domains	Rough sets	26	AA composition and physicochemical properties	87.9	91.3	97.1	86.0	90.8
	Component-coupling	15	AA composition	93.5	88.9	90.4	84.5	89.2
	Neural network	24	AA composition	86.0	96.0	88.2	86.0	89.2
	SVM	27	AA composition	88.8	95.2	96.3	91.5	93.2
	LogitBoost	30	AA composition	92.5	96.0	97.1	93.0	94.8
	SVM	This paper	PSI-BLAST based p-collocated AA pairs	98.0	93.3	95.6	93.4	94.9
	IB1	This paper	PSI-BLAST based p-collocated AA pairs	95.0	95.8	97.8	94.2	**95.7**
1189 domains (40% sequence identity)	Bayesian classifier	25	AA composition	54.8	57.1	75.2	22.2	53.8
	Logistic regression	32	AA composition, autocorrelations, and physicochemical properties	60.2	60.5	55.2	33.2	53.9
	SVM	32	AA composition, autocorrelations, and physicochemical properties	–	–	–	–	52.1
	Ensemble	28	AA composition, autocorrelations, and physicochemical properties	–	–	–	–	58.9
	SVM	This paper	PSI-BLAST based p-collocated AA pairs	75.8	75.2	82.6	31.8	**67.6**
	IB1	This paper	PSI-BLAST based p-collocated AA pairs	65.3	67.7	79.9	40.7	64.7
675 domains (30% sequence identity)	Bayesian classifier	25	AA composition	53.5	42.3	68.3	28.3	48.0
	SVM	This paper	PSI-BLAST based p-collocated AA pairs	74.3	59.6	79.7	34.5	**61.5**
	IB1	This paper	PSI-BLAST based p-collocated AA pairs	54.9	47.4	68.9	35.0	51.5

The best results are shown in bold.

sition-based sequence representation. Similarly, for the dataset with 204 domains, the proposed representation results in 1.0/6.9 = 14% error rate reduction when compared with the existing best-performing augmented covariant discriminate algorithm, which applies sequence representation that combines pseudo AA composition and dipeptide frequencies (total of 420 features).[43] Compared with this representation, our feature set includes only 50 features. For the dataset with 1189 domains that share 40% sequence identity, the proposed representation gives 8.7/41.1 = 21% error rate reduction when compared with best prior results reported for the ensemble classifier that uses custom-designed representation that includes AA composition, autocorrelations, and physicochemical properties. For dataset with 675 domains that share up to 30% sequence similarity, the proposed representation gives 13.5/52 = 26% error rate reduction compared with the Bayesian classifier. In short, the accuracies obtained by using the proposed sequence representation are substantially better than the best, previously reported accuracies for all five reported datasets. Finally, our method was also evaluated on the new dataset that includes 640 sequences with low 25% identity. For this dataset, the overall accuracy is 62.3%, see Table 2, which is comparable with the results on dataset with higher, 30% sequence similarity.

In contrast with the relatively high accuracies obtained by majority of existing methods on datasets with 204, 277, and 498 domains, the accuracies on other three datasets are lower. We believe that the high accuracies are an artefact resulting from duplicates and highly similar sequences included in these three datasets, which corroborates with results reported in ref. 32. For instance, the datasets with 498 domains include over 10 copies of the same sequence that appears under different PDB IDs.

Finally, we note that the reported best classifiers that use the proposed sequence representation are relatively simple, i.e., a single SVM and a nearest neighbor-based method, when compared with the best reported methods that include complex, ensemble-based classifiers, i.e., Logit Boost[18] and ensemble reported in ref. 28. Although this simplicity eases the implementation, an ensemble-based method that applies the proposed representation, which will constitute our future work, could provide further improvements.

Conclusions

Prediction of the protein structural class is an important and difficult problem. This paper focuses on the design of a high-quality sequence representation that allows improving the prediction accuracy when compared with the currently used representations and classification algorithms. We show that the proposed PSI-BLAST profile-based collocation of AA pairs is a novel and promising feature representation. Our comprehensive empirical tests that include five benchmark datasets show that the accuracy of the structural class prediction can be substantially improved by applying this representation, i.e., relatively simple classifiers that use the proposed features provide better accuracy than existing, best-performing, more complex classifiers that use other representations. The corresponding error rate reductions range between 14% and 26% over five test datasets considered in this study. The new representation can be extended to other protein prediction tasks such as fold, solvent accessibility, membrane protein type, and enzyme family predictions.

Finally, the structural class prediction still faces challenging issues such as relatively low accuracy for the $\alpha + \beta$ class, especially for datasets with low sequence identity. Also, as originally discussed,[32] the predictions for sequences with low similarity are characterized by lower quality. Our results show that for a dataset with 40% sequence identity, the prediction accuracy drops to 67%, while for datasets with even lower 30% and 25% identity, the accuracy further drops to about 62%.

References

1. Levitt, M.; Chothia, C. Nature 1976, 261, 552.
2. Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. J Mol Biol 1995, 247, 536.
3. Gromiha, M. M.; Selvaraj, S.; Thangakani, A. M. J. Chem Inf Model 2006, 46, 1503.
4. Kuznetsov, I. B.; Gou, Z.; Li, R.; Hwang, S. Proteins 2006, 64, 19.
5. Gromiha, M. M. Biophys Chem 2005, 117, 65.
6. Gromiha, M. M.; Suwa, M. Bioinformatics 2005, 21, 961.
7. Gromiha, M. M. J. Chem Inf Model 2005, 45, 494.
8. He, H.; McAllister, G.; Smith, T. F. Proteins 2002, 48, 654.
9. Zhang, C. T.; Zhang, Z.; He, Z. J Prot Chem 1998, 17, 261.
10. Zhang, Z.; Sun, Z. R.; Zhang, C. T. J Theor Biol 2001, 208, 65.
11. Chou, K. C. J Mol Biol 1992, 223, 509.
12. Carlacci, L.; Chou, K. C.; Maggiora, G. M. Biochemistry 1991, 30, 4389.
13. Klein, P.; Delisi, C. Biopolymers 1986, 25, 1659.
14. Nakashima, H.; Nishikawa, K.; Ooi, T. J. Biochem 1986, 99, 153.
15. Zhou, G. P. J Prot Chem 1998, 17, 729.
16. Bu, W.-S.; Feng, Z.-P.; Zhang, Z.; Zhang, C.-T. Eur J Biochem 1999, 266, 1043.
17. Jin, L.; Fang, W.; Tang, H. Comput Biol Chem 2003, 27, 373.
18. Cai, Y. D.; Feng, K. Y.; Lu, W. C.; Chou, K. C. J Theor Biol 2006, 238, 172.
19. Sun, X.-D.; Huang, R. B. Amino Acids 2006, 30, 469.
20. Xiao, X.; Shao, S.; Huang, Z.; Chou, K. C. J Comput Chem 2006, 27, 478.
21. Zhang, C. T.; Chou, K. C. Prot Sci 1992, 1, 401.
22. Chou, K. C.; Zhang, C. T. J Prot Chem 1993, 12, 169.
23. Shen, H. B.; Yang, J.; Liu, X.-J.; Chou, K. C. Biochem Biophys Res Commun 2005, 334, 577.
24. Cai, Y.; Zhou, G. Biochimie 2002, 82, 783.
25. Wang, Z.-X.; Yuan, Z. Proteins 2000, 38, 165.
26. Cao, Y.; Liu, S.; Zhang, L.; Qin, J.; Wang, J.; Tang, K. BMC Bioinf 2006, 7, 20.
27. Cai, Y. D.; Liu, X. J.; Xu, X.; Zhou, G. P. BMC Bioinf 2001, 2, 3.
28. Kedarisetti, K. D.; Kurgan, L.; Dick, S. Biochem Biophys Res Commun 2006, 348, 981.
29. Dong, L.; Yuan, Y.; Cai, Y. J Biomol Struct Dyn 2006, 24, 239.
30. Feng, K. Y.; Cai, Y. D.; Chou, K. C. Biochem Biophys Res Commun 2005, 334, 213.
31. Chou, K. C. Curr Prot Pept Sci 2005, 6, 423.
32. Kurgan, L.; Homaeian, L. Pattern Recogn 2006, 39, 2323.
33. Kurgan, L.; Chen, K. Biochem Biophys Res Commun 2007, 357, 453.
34. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J. H.; Zhang, Z.; Miller, W.; Lipman, D. J. Nucleic Acid Res 1997, 17, 3389.
35. Jones, D. T. J Mol Biol 1995, 292, 195.
36. Kim, H.; Park, H. Proteins 2004, 54, 557.

37. Jones, D. T. Bioinformatics 2007, 23, 538.
38. Chen, K.; Kurgan, L.; Rahbari, M. Biochem Biophys Res Commun 2007, 355, 764.
39. Chen, K.; Kurgan, L. A.; Ruan, J. BMC Struct Biol 2007, 7, 25.
40. Chou, K. C. Biochem Biophys Res Commun 1999, 264, 216.
41. Zhang, C. T.; Chou, K. C.; Maggiora, G. M. Prot Eng 1995, 8, 425.
42. Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Comput Chem 2002, 26, 293.
43. Lin, H.; Li, Q. Z. J Comput Chem 2007, 28, 1463.
44. Chen, C.; Tian, Y. X.; Zou, X. Y.; Cai, P. X.; Mo, J. Y. J Theor Biol 2006, 243, 444.
45. Zhang, T. L.; Ding, Y. S. Amino Acids 2007, 33, 623.
46. Smith, T. F.; Waterman, M. S. J Mol Biol 1981, 147, 195.
47. Chen, K.; Kurgan, L.; Ruan, J. IEEE Symp Comp Intell Bioinf Comp Biol 2006, 366.
48. Witten, I.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques; Morgan Kaufmann: San Francisco, 2005.
49. Keerthi, S. S.; Shevade, S. K.; Bhattacharyya, C.; Murphy, K. R. K. Neur Comput 2001, 13, 637.
50. Le, C. S.; Houwelingen, J. C. Appl Stat 1992, 41, 191.
51. Aha, D.; Kibler, D. Machine Learn 1991, 6, 37.
52. John, G. H.; Langley, Proc 11th Conf Uncertainty Artific Intell 1995, 338.
53. Ross, Q. C4.5: Programs for Machine Learning; Morgan Kaufmann Publishers: San Mateo, CA, 1993.