Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 357 (2007) 453–460
BBRC
www.elsevier.com/locate/ybbrc

Prediction of protein structural class for the twilight zone sequences

Lukasz Kurgan *, Ke Chen

Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alta., Canada

Received 21 March 2007
Available online 5 April 2007

Abstract

Structural class characterizes the overall folding type of a protein or its domain. This paper develops an accurate method for in-silico prediction of structural classes from low homology (twilight zone) protein sequences. The proposed LLSC-PRED method applies linear logistic regression classifier and a custom-designed, feature-based sequence representation to provide predictions. The main advantages of the LLSC-PRED are the comprehensive representation that includes 58 features describing composition and physicochemical properties of the sequences and transparency of the prediction model. The representation also includes predicted secondary structure content, thus for the first time exploring synergy between these two related predictions. Based on tests performed with a large set of 1673 twilight zone domains, the LLSC-PRED's prediction accuracy, which equals over 62%, is shown to be better than accuracy of over a dozen recently published competing in silico methods and similar to accuracy of other, non-transparent classifiers that use the proposed representation.

Keywords: SCOP structural class; Structural class prediction; Secondary structure; Twilight zone proteins; Low sequence homology; Sequence representation

Despite assuming a multitude of complex three-dimensional structures and bearing a wide range of biological functions proteins are characterized by simple and regular local folding patterns. Structural class categorizes various proteins into groups that share similarities in the local folding. Prediction of structural classes is based on identifying these folding patterns based on thousands of already categorized proteins, and applying these patters to millions of proteins with unknown structures but known amino acid (AA) sequences, i.e., as of March 13, 2007, release 22 of the NCBI's RefSeq database stores 3,438,099 sequences.

One of the most accurate classifications of structural classes can be found in the expert-curated SCOP (Structural Classification of Proteins) database [32] (as of October 2006, release 1.71 of SCOP stores 75,930 sequences). The basic structural unit of classification in this database is either the entire sequence or a protein domain (structurally conserved fragment of the sequence). SCOP database is organized as a hierarchy of known protein and protein domain structures where first level is based on the structural class. There are four major structural classes: all-α, all-β, α/β, and α + β. The all-α and all-β classes represent structures that consist of mainly α-helices and β-strands, respectively. The α/β and α + β classes contain both α-helices and β-strands where the α/β class includes mainly parallel β-strands and α + β class includes anti-parallel strands.

Structural class prediction is usually performed in two steps. First, the AA sequences are transformed into a fixed-length feature vectors. Next, the feature vectors are fed to a classification algorithm to perform the prediction. The last twenty years have seen numerous methods for computational prediction of protein structural class. Majority of the prediction methods use relatively simple sequence representations such as composition vectors, auto-correlation function based on non-bonded residue

* Corresponding author. Fax: +1 780 492 1811.
E-mail address: lkurgan@ece.ualberta.ca (L. Kurgan).

0006-291X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2007.03.164

energy, polypeptide composition, pseudo AA composition and complexity measure factor [2,7,22,26,41,46]. A few recent methods developed more advanced representations that either combine physicochemical properties and sequence composition or optimize one selected type of the representation [21,23,25]. Different classification algorithms, including fuzzy clustering [38], neural network [4], Bayesian classification [39], rough sets [8], component-coupled [46], information discrepancy [22,26], logistic regression [21,25,23], decision tree [7,10], and support vector machine [5,6,10,25], have been already used. Recent works also explored application of complex classification models, such as ensembles [25], bagging [10], and boosting [7,14]. However, some of these algorithms were tested on small datasets with uncontrolled (often high) sequence homology, which results in an overestimated prediction accuracy [23]. At the same time, secondary structure of homologous sequences can be reliably predicted [34], and this information can be used to come up with the corresponding structural class.

On the other hand, low homology sequences pose a substantial challenge. Virtually all secondary structure prediction methods use sequence alignment that requires at least ~30% homology between the query sequence and sequence(s) used to predict its structure [37]. The proteins characterized by a lower, 20–30%, homology with sequences that are used to predict their structure are called the twilight zone proteins [35]. More than 95% of all sequence pairs detected in the twilight zone have different structures [35], which significantly reduces the prediction accuracy. For instance, prediction of the secondary structure for homologous sequences by the state-of-the-art alignment-based methods yields about 80% accuracy, while for the twilight zone sequences it drops to only 65–68% [31]. Similarly, in case of structural class prediction accuracies for highly homologous protein datasets reach over 90%, while they drop to about 57% in case of the twilight zone sequences [23].

A substantial number of sequences for which tertiary structure was recently solved belong to the twilight zone, demonstrating the extent of the problem. We collected all sequences entered to the Protein Data Bank (PDB) in 2005 and aligned these sequences (using Smith–Waterman algorithm) with sequences stored in PDB before 2005. Among 1657 sequences, 40.1% belong to the twilight zone, 27% belong to non-twilight zone, and 32.9% were identical (100% sequence homology) when compared with proteins stored in PDB before 2005. The large number of twilight zone proteins that are of interest to the community and relatively low prediction accuracy for these sequences that is provided by the existing structural class prediction methods serve as our motivation.

To this end, we propose a novel in silico method that aims to improve prediction accuracy for the twilight zone proteins. The proposed method, referred to as LLSC-PRED, uses a custom-designed sequence representation and a transparent linear logistic regression model to predict structural classes.

Materials and methods

Dataset. The proposed method is designed and tested on a large set of twilight zone sequences. The dataset, referred to as 25PDB, was selected based on the 25% PDBSELECT list [18], which includes proteins scanned with high resolution and with low, on average 25%, homology. The dataset was originally developed and published in [23]. It contains 1673 proteins and domains, which include 443 all-α, 443 all-β, 346 α/β, and 441 α + β sequences.

Feature based sequence representation. The sequence representation design is based on a comprehensive list of feature sets that were previously used for prediction of protein structural class, secondary structure content, function, family, and solvent accessibility, and a set of four new features that correspond to the predicted secondary structure content, see Tables 1 and 2. The content prediction aims to quantify the amount of residues in the sequence that assume helical and strand conformation. The beneficial impact of the predicted content was first investigated in [24]. The LLSC-PRED method is the first to apply the content prediction to improve the structural class prediction. Two recent content prediction methods were used [29,45] and the strand/helix content values were computed based on protein sequences and using 10-fold cross validation. The resulting set of 2121 features was next processed by a feature selection method to obtain the final, customized representation.

Proposed sequence representation. Since the considered feature sets include large number of features, a feature selection method was used to reduce the dimensionality and potentially improve the prediction accuracy. A correlation-based feature subset selection method (CFSS) was used [17]. It evaluates a given subset of features, which is found using best first search based on the hill-climbing with backtracking, by considering the individual predictive ability of each feature along with the degree of redundancy between them. We also tested half a dozen other feature selection methods, and concluded that the CFSS provides the best results. The feature selection was performed using 10-fold cross validation to avoid overfitting, and features that were found significant by the CFSS in at least 1-fold were selected. As a result, 58 features were selected among the original set of 2021 features, see Table 3.

The proposed representation includes at least one feature from each feature set listed in Table 1; we also tried other feature sets but only those that provide significant features are listed. The most consistent, i.e., selected in all folds, features include the predicted content values (for both secondary structures and prediction methods) and autocorrelation functions. The composition-based features, with significant number of features related to collocated AAs pairs, were also found useful.

Prediction method. The structural class prediction was performed using linear logistic regression classifier, which is based on LogitBoost learning with linear regression functions as base learners [28]. In this classifier, a given instance (protein sequence) is classified into one of the four structural classes using

$$j_{\text{predicted}} = \arg\max_j P(C = j \mid X = x)$$

where class variable C takes on four values $j = \{\text{all-}\alpha, \text{all-}\beta, \alpha/\beta, \alpha+\beta\}$, $P(C = j \mid X = x)$ is to the posterior class probability for class j, and x is the feature vector that represents the sequence. Logistic regression models posterior probabilities using linear functions in x ensuring that they sum to 1 and remain in $[0,1]$. The linear regression model is specified in terms of log-odds that separate each class from the "base class" J such that

$$\log \frac{P(C = j \mid X = x)}{P(C = J \mid X = x)} = \xi_{0,j} + \xi_j^T x$$

where the parameter vector ξ_j is estimated (learned) based on training sequences (in their feature-based representation) using LogitBoost algorithm [15]. As a result, the prediction model is transparent, i.e., ξ_j values can be used to estimate the relative predictive value of the corresponding features.

Table 1
Feature sets used to develop the proposed sequence representation

Feature sets (# features)		Description	Prior applications	References
Sequence length (1)		N (# AA in the sequence)	Secondary structure content, and structural class predictions	[19,23,25,33]
Index-based	Average isoelectric point (1)	$pI = \frac{1}{N}\sum_{i=1}^{N} pI_i$, see Table 2 for pI_i values		
	Auto-correlation functions based on FH_i, EH_i and Hp indices (25)	$A_n^a = \frac{1}{N-n}\sum_{i=1}^{N-n} a_i a_{i+n}$ where a defines the corresponding physicochemical AA index; two hydrophobicity indices, i.e., the Fauchere–Pliska's (FH) with $n = 1,2,\ldots,10$ [13] and the Eisenberg's (EH) [11] $n = 1,2,\ldots,6$, and the hydropathy (Hp) index [27] with $n = 1,2\ldots,9$ were used; see Table 2	Protein content, structural class, and solvent accessibility prediction	[23,25,29,30,44,45]
	Auto-correlation functions based on cumulative FH_i index (6)	$\mathrm{Acum}_n^a = \frac{\sum_{i=1}^{N-n}\left(\sum_{j=1}^{i} a_j\right)\times\left(\sum_{j=1}^{i+n} a_j\right)}{N-n}$ where a is the Fauchere–Pliska's (FH) [13] index with $n = 1,2,\ldots,6$; see Table 2	Protein content prediction	[20]
	Sum of hydrophobicities based on FH_i and EH_i indices (2)	$H_{\mathrm{sum}}^a = \sum_{i=1}^{N} a_i$ where a is the Fauchere–Pliska's (FH) [13] or the Eisenberg's (EH) [11] index; see Table 2	Protein structural class prediction	[25]
	Sum of 3-running average of hydrophobicities based on FH_i and EH_i indices (2)	$H_{\mathrm{sum3}}^a = \sum_{i=1}^{N-3}\left(\sum_{j=i}^{i+3} a_j\right)\Big/3$ where a is the Fauchere–Pliska's (FH) [13] or the Eisenberg's (EH) [11] index; see Table 2	Protein structural class prediction	[25]
AA composition	Composition vector (20)	CV_i the composition percentage of ith AA in the sequence; see Table 2	Protein structure, structural class and content predictions	[7–9,12,23,25,33,36,43–45]
	Composition of collocated AA pairs (2000)	$CV_{AA_iAA_j}$, $CV_{AA_i\text{-}AA_j}$, $CV_{AA_i\text{--}AA_j}$, $CV_{AA_i\text{---}AA_j}$, $CV_{AA_i\text{----}AA_j}$ for ith and jth AAs, see Table 2 for the AA index assignment; this is the composition percentage of collocated AA pairs, i.e., AA_iAA_j denotes dipeptides, AA_i-AA_j, denotes two AAs separated by a single gap, and AA_i--AA_j, AA_i---AA_j, and AA_i----AA_j denote two AAs separated by 2, 3 and 4 gaps, respectively; there are 400 pairs for each gap size	Secondary structure prediction	[3]
	First and second order composition moment vector (40)	$CMV_i^k = \frac{\sum_{i5j=1}^{x_i} n_{ij}^k}{\prod_{d=1}^{k}(N-k)}$ where n_{ij} represents the jth position of the ith AA, n_i is the frequency of ith AA in the sequence, and $k = 1,2$ is the order of the CMV; CMV for $k = 0$ reduces to CV	Protein content and structural class predictions	[23,25,36]
Property groups	R groups (5)	RG_i where $i = 1,2,\ldots,5$; $i = 1$ corresponds to nonpolar aliphatic AAs (AVLIMG), $i = 2$ to polar uncharged AAs (SPTCNQ), $i = 3$ to positively charged AAs (KHR), $i = 4$ to negative AAs (DE), and $i = 5$ to aromatic AAs (FYW); the composition percentage of each group in the sequence is computed	Protein structural class and secondary structure predictions	[25,42]
	Electronic groups (5)	EG_i where $i = 1,2,\ldots,5$; $i = 1$ corresponds to electron donor AAs (DEPA), $i = 2$ to weak electron donor AAs (LIV), $i = 3$ to electron acceptor AAs (KNR), $i = 4$ to weak electron acceptor AAs (FYMTQ), and $i = 5$ to neutral AAs (GHWS); the composition percentage of each group in the sequence is computed	Protein secondary structure and structural class predictions	[16,23,25]
	Chemical groups (10)	CG_i these groups are defined based on composition of chemical group that constitute the side chains [16] where $i = 1,2,\ldots,10$ corresponds to C, CAROM, CH, CH_2, CH_2RING, CH_3, CHAROM, CO, NH, OH side chain groups, respectively. The composition percentage of each chemical group in all side chains in the sequence is computed		
Secondary structure content (4)		$\mathrm{Content}H_f$ and $\mathrm{content}E_f$ where H corresponds to helix content, E corresponds to strand content and f corresponds to the prediction method, i.e., method by Lin and Pan (LP) [30] and by Zhang and colleagues (Z) [45]; the content values were predicted using 10-fold cross validation	Protein content prediction	[30,45]

Table 2
The AA indices

Amino acid	Code	AA index	Physicochemical index			
			p*I*	FH	EH	Hp
Alanine	A	1	6.01	0.42	0.62	1.8
Cysteine	C	2	5.07	1.34	0.29	2.5
Aspartate	D	3	2.77	−1.05	−0.9	−3.5
Glutamate	E	4	3.22	−0.87	−0.74	−3.5
Phenylalanine	F	5	5.48	2.44	1.19	2.8
Glycine	G	6	5.97	0	0.48	−0.4
Histidine	H	7	7.59	0.18	−0.4	−3.2
Isoleucine	I	8	6.02	2.46	1.38	4.5
Lysine	K	9	9.74	−1.35	−1.5	−3.9
Leucine	L	10	5.98	2.32	1.06	3.8
Methionine	M	11	5.47	1.68	0.64	1.9
Asparagine	N	12	5.41	−0.82	−0.78	−3.5
Proline	P	13	6.48	0.98	0.12	−1.6
Glutamine	Q	14	5.65	−0.3	−0.85	−3.5
Arginine	R	15	10.76	−1.37	−2.53	−4.5
Serine	S	16	5.68	−0.05	−0.18	−0.8
Threonine	T	17	5.87	0.35	−0.05	−0.7
Valine	V	18	5.97	1.66	1.08	4.2
Tryptophan	W	19	5.89	3.07	0.81	−0.9
Tyrosine	Y	20	5.67	1.31	0.26	−1.3

Result and discussion

Experimental setup

Classification algorithms used to develop and compare the proposed method were implemented in Weka, which is a comprehensive open-source library of machine learning methods [40]. The proposed LLSC-PRED method was comprehensively compared with competing methods which use other sequence representations and best performing Support Vector Machine (SVM) classifier using the same 58 features to represent sequences. The comparison includes three state-of-the-art groups of competing algorithms:

- methods that apply optimized representations [21,23,25]
- recent advanced multi-classifier systems including boosting [7], ensembles [25], and bagging [10]
- best performing SVM [6] and information discrepancy based algorithms [22,26].

The experimental evaluation was performed using two out-of-sample tests, i.e., 10-fold cross validation and jackknife tests, to avoid overfitting and assure statistical validity of the results [23]. The tests were performed on the twilight zone 25PDB dataset, and the reported results include *overall accuracy* (the number of correct predictions divided by the total number of test sequences), *accuracy for each structural class* (number of correct predictions for a given class divided by the number of sequences from this class), *Matthews's correlation coefficient* (MCC) for each structural class, and *generalized squared correlation* (GC^2). The MCC values range between −1 and 1, where 0 represents random correlation, and bigger positive (negative) values indicate better (lower) quality of the prediction for a given structural class. Since MCC works only for binary classification, we also reported GC^2, which is based on χ^2 statistics. The GC^2 values range between 0 and 1, where 0 corresponds to worst possible classification (no correct predictions) and 1 corresponds to perfect classification. MCC and GC^2 are described in detail in [1].

Results and comparison with competing methods

The classification results for the 13 competing methods, LLSC-PRED and two SVM classifiers are compared in Table 4. The LLSC-PRED and the two SVMs use the proposed 58 features and were optimized to maximize overall accuracy based on 10-fold cross validation (we used both polynomial and Gaussian kernels for SVMs). The competing methods use the original author's setup including the sequence representation and the algorithm's parameters.

The LLSC-PRED gives over 62% accuracy for both out-of-sample tests. The only other comparable results are generated by using SVM on the proposed representation. Although LLSC-PRED and SVM share similar accuracy, the proposed linear logistic regression model is transparent and easy to interpret (see next section), while the SVM models are virtually impossible to comprehend. The remaining, competing methods obtain accuracies that range between 35% and 60%.

The only two competing methods that reach 60% accuracy are also based on a custom-designed representation that includes both composition and physicochemical prop-

Table 3
Proposed feature-based sequence representation

Features		# Folds, in which CFSS found a given feature significant
#	Feature names	
7	A^{Hp}_2, A^{Hp}_4, A^{FH}_2, contentH_{LP}, contentE_{LP}, contentH_Z, contentE_Z	10
2	p*I*, H^{FH}_{sum3}	8
1	CV_F	7
2	$CV_{D\text{-}V}$, EG_5	6
1	$CV_{L\text{---}G}$	5
5	CV_{EQ}, CV_A, CV_T, CG_{10}, H^{EH}_{sum}	4
8	$CV_{V\text{-}I}$, $CV_{V\text{-}V}$, $CV_{C\text{-}C}$, $CV_{V\text{---}G}$, CV_H, CV_P, CMV^2_Q, RG_3	3
12	$CV_{M\text{-}K}$, $CV_{V\text{--}G}$, $CV_{A\text{---}A}$, $CV_{G\text{---}V}$, $CV_{P\text{---}C}$, $CV_{V\text{---}P}$, N, CV_K, H^{FH}_{sum}, $Acum^{FH}_3$, $Acum^{FH}_6$, A^{Hp}_3	2
20	CV_{EY}, CV_{GI}, CV_{MN}, $CV_{C\text{-}Q}$, $CV_{I\text{-}I}$, $CV_{M\text{--}P}$, $CV_{W\text{--}N}$, $CV_{Y\text{--}C}$, $CV_{W\text{---}P}$, $CV_{G\text{----}V}$, $CV_{I\text{----}E}$, $CV_{I\text{----}G}$, CMV^1_Q, CMV^1_T, CMV^2_G, RG_1, EG_1, A^{EH}_2, H^{EH}_{sum3}, $Acum^{FH}_4$	1

Table 4
Summary of the experimental results; [22] was not originally tested using 10-fold cross validation and thus we also did not report these results

		Algorithm	Sequence representation (# features)	References	Accuracy					MCC				GC^2
					all-α	all-β	α/β	α + β	Overall	all-α	all-β	α/β	α + β	
Jackknife	Competing methods	SVM (Gaussian kernel)	CV (20)	[6]	68.6	59.6	59.8	28.6	53.9	0.52	0.42	0.43	0.15	0.17
		LogicBoost with decision tree	CV (20)	[7]	56.9	51.5	45.4	30.2	46.0	0.41	0.32	0.32	0.06	0.10
		Bagging with random tree	CV (20)	[10]	58.7	47.0	35.5	24.7	41.8	0.33	0.26	0.22	0.06	0.06
		LogitBoost with decision stump	CV (20)	[10]	62.8	52.6	50.0	32.4	49.4	0.49	0.35	0.34	0.11	0.13
		SVM (3rd order polyn. kernel)	CV (20)	[10]	61.2	53.5	57.2	27.7	49.5	0.46	0.35	0.39	0.11	0.13
		Multinomial logistic regression	Custom dipeptides (16)	[21]	56.2	44.5	41.3	18.8	40.2	0.23	0.20	0.31	0.06	0.05
		Information discrepancy	Dipeptides (400)	[22,26]	59.6	54.2	47.1	23.5	47.0	0.46	0.40	0.24	0.04	0.12
		Information discrepancy	Tripeptides (8000)	[22,26]	45.8	48.5	51.7	32.5	44.7	0.39	0.39	0.25	0.06	0.11
		Multinomial logistic regression	Custom (34)	[25]	71.1	65.3	66.5	37.3	60.0	0.61	0.51	0.51	0.22	0.25
		SVM with RBF kernel	Custom (34)	[25]	69.7	62.1	67.1	39.3	59.5	0.60	0.50	0.53	0.21	0.25
		StackingC ensemble	Custom (34)	[25]	74.6	67.9	70.2	32.4	61.3	0.62	0.53	0.55	0.22	0.26
		Multinomial logistic regression	Custom (66)	[23]	69.1	61.6	60.1	38.3	57.1	0.56	0.44	0.48	0.21	0.21
		SVM (1st order polyn. kernel)	Autocorrelation (30)	[23]	50.1	49.4	28.8	29.5	34.2	0.16	0.16	0.05	0.05	0.02
	This paper	SVM (1st order polyn. kernel)	Custom (58)	This paper	77.4	66.4	61.3	45.4	62.7	0.65	0.54	0.55	0.27	0.28
		SVM (Gaussian kernel)	Custom (58)	This paper	76.5	64.6	63.3	44.9	62.3	0.65	0.53	0.54	0.26	0.28
		LLSC-PRED	Custom (58)	This paper	75.2	67.5	62.1	44.0	62.2	0.63	0.54	0.54	0.27	0.27
10-fold cross validation	Competing methods	SVM (Gaussian kernel)	CV (20)	[6]	67.9	59.1	58.1	27.7	53.0	0.51	0.42	0.41	0.14	0.16
		LogicBoost with decision tree	CV (20)	[7]	51.9	53.7	46.5	32.4	46.1	0.38	0.37	0.31	0.07	0.10
		Bagging with random tree	CV (20)	[10]	53.5	51.0	37.6	22.0	41.2	0.28	0.30	0.22	0.04	0.06
		LogitBoost with decision stump	CV (20)	[10]	63.2	53.5	50.9	32.4	50.0	0.48	0.36	0.36	0.12	0.14
		SVM (3rd order polyn. kernel)	CV (20)	[10]	61.4	54.0	55.2	27.4	49.2	0.46	0.35	0.37	0.10	0.13
		Multinomial logistic regression	Custom dipeptides (16)	[21]	56.9	44.2	42.2	17.7	40.2	0.24	0.20	0.32	0.04	0.06
		Multinomial logistic regression	Custom (34)	[25]	69.9	65.3	66.5	38.4	60.0	0.60	0.52	0.51	0.23	0.25
		SVM with RBF kernel	Custom (34)	[25]	70.2	61.6	67.6	39.6	59.8	0.60	0.49	0.53	0.22	0.25
		StackingC ensemble	Custom (34)	[25]	73.4	67.3	69.1	29.8	59.9	0.59	0.52	0.54	0.18	0.25
		Multinomial logistic regression	Custom (66)	[23]	69.1	60.5	59.5	38.1	56.7	0.56	0.44	0.48	0.20	0.21
		SVM (1st order polyn. kernel)	Autocorrelation (30)	[23]	52.4	49.7	0.3	30.4	35.1	0.18	0.16	0.05	0.06	0.02
	This paper	SVM (1st order polyn. kernel)	Custom (58)	This paper	77.7	66.8	60.7	45.4	62.8	0.64	0.54	0.54	0.28	0.28
		SVM (Gaussian kernel)	Custom (58)	This paper	76.1	65.0	63.6	45.8	62.6	0.66	0.53	0.54	0.27	0.28
		LLSC-PRED	Custom (58)	This paper	74.7	66.4	62.7	45.8	62.4	0.63	0.54	0.54	0.27	0.28

erties [25]. The most accurate method that uses the most popular and simple composition vector based representation obtained 54% accuracy [6]. In general, simple representations result in low accuracy for the twilight zone proteins. This highlights the importance of the sequence representation, i.e., we believe that future improvements will be possible by designing more advanced representations rather than using more advanced classification methods. On the other hand, SVM and logistic repression classifiers are shown to perform the best on this challenging, low-homology prediction problem.

The most accurate predictions concern all-α class (75–77% accuracy), while the best results for all-β and α/β classes range between 65% and 67% and between 61% and 63%, respectively. The lowest accuracy (44–45%) is obtained for the α + β class. This trend is universal for all tested method, although the corresponding accuracies are lower. The main reason for good performance for all-α class is that these sequences are helix rich and helical structures are the easiest to predict, i.e., a helix is formed by a single, continuous sequence segment and is characterized by highly repetitive structure.

Finally, we note that evaluations using both accuracy/class accuracy and GC^2/class MCC result in very similar conclusions.

Analysis of the prediction model

The LLSC-PRED generated a linear prediction model shown in Table 5. The remaining features, i.e., those among the selected 58 features that are not listed in Table 5, assumed corresponding parameter vector values equal 0 for all four structural classes.

The model shows relative predictive value of individual features for each of the structural classes. The features with negative (positive) ξ_j values have detrimental (promoting) effect on the prediction, i.e., their positive values result in lower (higher) score while the structural class with the highest score is selected.

For the all-α class, the predicted strand content has strong detrimental effect; this correlates with the class definition. A negative relation between composition moment of Glutamine (CMV^2_Q) and promoting effect of the Proline's composition (CV_P) and hydrophobicity-based autocorrelation ($A^{EH}{}_2$) are also characteristic for this class.

For the all-β class, the predicted strand (helix) content has strong promoting (inhibitory) effect; this again agrees with the class definition. The Histidine's composition (CV_H) is found to be positively associated with this class, while the increased amount of Alanine (CV_A) lowers the probability of the corresponding sequence to belong to the all-β class.

For the α/β class, the strongest promoting effect is associated with the composition moment of Glutamine (CMV^2_Q) and the high amount of nonpolar aliphatic residues in the sequence (RG_1). The inhibitory effect is pro-

Table 5
The LLSC-PRED's prediction model for the 25PDB dataset; strongest contributing features for each structural class are shown in bold and the features are ordered by their average (over the four classes) values

	CMV^2_Q	$ContentE_Z$	CV_P	$ContentH_Z$	RG_3	RG_1	CV_H	CV_A	CMV^2_G	$ContentE_{LP}$	CV_T	CG_{10}	$A^{EH}{}_2$	CV_K	CV_F	EG_1	CMV^1_Q	CMV^1_T	CV_{C-C}	$A^{FH}{}_2$	$A^{Hp}{}_2$	CV_{MN}	CV_{C-Q}
$\xi_{all-\alpha}$	**−26.2**	**−13.73**	**12.56**	0	0	0	−2.95	2.96	0	**−7.78**	0	0	**5.18**	0	−3.42	0	0	0	.26	−.32	−.5	0	.16
$\xi_{all-\beta}$	0	**4.06**	0	**−12.15**	3.58	1.57	**4.35**	**−6.46**	0	0	2.97	3.65	−1.03	0	0	0	0	0	−1.36	0	.71	−.38	0
$\xi_{\alpha/\beta}$	**11.42**	0	−2.87	0	**−4.68**	**9.43**	0	0	**−5.25**	0	−3.78	2.79	0	1.96	0	2.33	0	0	0	0	0	.49	0
$\xi_{\alpha+\beta}$	0	0	0	−2.13	**−4.28**	0	**3.22**	0	**4.37**	1.49	0	0	−.33	**−4.6**	0	0	1.57	1.53	0	.95	.18	.14	−1.09
|Average|	9.4	4.4	3.9	3.6	3.1	2.8	2.6	2.4	2.4	2.3	1.7	1.6	1.6	1.6	0.9	0.6	0.4	0.4	0.4	0.3	0.3	0.3	0.3

	$A^{Hp}{}_4$	CV_{GI}	CV_{EY}	CV_{I-I}	CV_{M-K}	CV_{V-V}	CV_{Y-C}	CV_{W-N}	CV_{V-P}	$A^{Hp}{}_3$	CV_{EQ}	CV_{D-V}	CV_{V-I}	CV_{M-P}	CV_{V-G}	CV_{G---V}	CV_{A--A}	CV_{P--C}	CV_{W--P}	CV_{L---G}	CV_{I---G}	CV_{G---V}	$H^{FH}{}_{sum3}$	N (length)	CV_{I----E}	CV_{V---G}	ξ_0
$\xi_{all-\alpha}$	.14	.1	−.2	−.37	.43	−.48	0	0	−.45	0	.23	0	−.37	−.27	−.24	−.22	0	0	−.33	0	−.3	−.18	0	0	0	−.07	4.39
$\xi_{all-\beta}$	−.36	−.21	0	−.09	−.28	0	−.26	−.14	.28	−.05	−.27	0	0	0	0	−.09	0	0	0	0	0	0	0	0	−.07	.05	.72
$\xi_{\alpha/\beta}$	−.28	.36	.44	.28	0	.19	.44	0	0	.09	0	.12	.06	.11	.06	.13	0	0	.15	.22	.11	0	.02	−.01	.06	0	−5.54
$\xi_{\alpha+\beta}$	.14	.04	0	0	−.05	−.1	0	−.58	0	−.25	−.04	0	0	.07	−.05	0	−.01	.3	0	−.07	−.13	−.03	−.01	0	0	0	2.71
|Average|	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	3.3

vided by the positively charged AAs (RG_3) and the composition moment of Glycine (CMV^2_G).

The $\alpha + \beta$ class is characterized by the lowest ξ_j values, which results in the lowest prediction accuracy. The only significant features include Histidine's composition (CV_H), composition moment of Glycine (CMV^2_G), and the detrimental effect of the large number of positively charged AAs (RG_3) and Lysine residues (CV_K).

Individual features usually provide promoting effect on one class, while at the same time they have detrimental effect on other classes. In other words, they are selective for a given class. For instance, (CMV^2_Q) has positive impact on the α/β class and detrimental effect on all-α class, predicted strand content has detrimental effect on all-α class and positive effect on all-β class, RG_3 has promoting effect on all-β class and inhibitory effect on both α/β and $\alpha + \beta$ classes, etc. Finally, we note that RG_3 and RG_1 groups and CV_H and CV_A compositions were also included in the sequence representation proposed in [25].

Summary and conclusions

The structural class prediction for the twilight zone sequences is a challenging problem. This paper presents a novel approach that aims to improve the prediction accuracy via designing a composite (of AA composition, physicochemical properties and predicted secondary structure content) sequence representation. In contrast to equally well performing Support Vector Machine (SVM) based classifier, the proposed LLSC-PRED method applies easy to comprehend and fast to train linear logistic regression classifier.

Based on an extensive experimental comparison between the proposed and over a dozen of competing methods, the LLSC-PRED is shown to break 60% barrier [39] and achieves overall accuracy of over 62%. This result is matched only by SVM that applies the proposed sequence representation. To compare, the accuracy of the best-performing SVM that applies a simple, non-composite representation equals 54%.

The main contribution of this paper is the proposed sequence representation that includes 58 features. This representation, together with the transparent prediction model may help uncovering hidden relations between important physicochemical sequence properties and the structural classes. Based on our experimental results, we believe that future improvements will be possible by designing better sequence representations rather than applying more complex classifiers.

References

[1] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424.

[2] W.-S. Bu, Z.-P. Feng, Z. Zhang, C.-T. Zhang, Prediction of protein (domain) structural classes based on amino-acid index, European Journal of Biochemistry 266 (1999) 1043–1049.

[3] K. Chen, L. Kurgan, J. Ruan, Optimization of the sliding window size for protein structure prediction. 2006 International Conference on Computational Intelligence in Bioinformatics and Computational Biology (2006) 366–372.

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

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

[6] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Support vector machines for prediction of protein domain structural class, Journal of Theoretical Biology 221 (1) (2003) 115–120.

[7] Y.D. Cai, K.Y. Feng, W.C. Lu, K.C. Chou, Using LogitBoost classifier to predict protein structural classes, Journal of Theoretical Biology 238 (1) (2006) 172–176.

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

[9] K.C. Chou, Y.D. Cai, Prediction protein structural class by functional domain composition, Biochemical and Biophysical Research Communications 321 (2004) 1007–1009.

[10] L. Dong, Y. Yuan, T. Cai, Using bagging classifier to predict protein domain structural class, Journal of Biomolecular Structure & Dynanmics 24 (3) (2006) 239–242.

[11] D. Eisenberg, R.M. Weiss, T.C. Trewilliger, The hydrophobic moment detects periodicity in protein hydrophobicity, Proceedings of the National Academy of Science 81 (1) (1984) 140–144.

[12] F. Eisenhaber, F. Imperiale, P. Argos, C. Frommel, Pediction of secondary structural contents of proteins from their amino acid composition alone, I new analytic vector decomposition methods, Proteins 25 (2) (1996) 157–168.

[13] J.L. Fauchere, V. Pliska, Hydrophobic parameters *p* of amino-acid side chains from the partitioning of *N*-acetyl-amino-acid amides, European Journal of Medicinal Chemistry 18 (1983) 369–375.

[14] K.Y. Feng, Y.D. Cai, K.C. Chou, Boosting classifier for predicting protein domain structural class, Biochemical and Biophysical Research Communications 334 (1) (2005) 213–217.

[15] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Annals of Statistic 38 (2) (2000) 337–374.

[16] M.K. Ganapathiraju, J. Klein-Seetharaman, N. Balakrishnan, R. Reddy, Characterization of protein secondary structure, IEEE Signal Processing Magazine 15 (2004) 78–87.

[17] M. Hall, Correlation based feature selection for machine learning, Ph.D. dissertation, University of Waikato, Department of Computer Science, 1999.

[18] U. Hobohm, C. Sander, Enlarged representative set of protein structures, Protein Science 3 (1994) 522.

[19] U. Hobohm, C. Sander, A sequence property approach to searching protein databases, Journal of Molecular Biology 251 (1995) 390–399.

[20] L. Homaeian, Towards improving accuracy of protein content prediction for low homology sequences, M.Sc. thesis, University of Alberta, Department of Electrical and Computer Engineering, 2006.

[21] S. Jahandideh, P. Abdolmaleki, M. Jahandideh, S.H. Sadat Hayatshahi, Novel hybrid method for the evaluation of parameters contributing in determination of protein structural classes, Journal of Theoretical Biology 244 (2) (2007) 275–281.

[22] L. Jin, W. Fang, H. Tang, Prediction of protein structural classes by a new measure of information discrepancy, Computational Biology and Chemistry 27 (2003) 373–380.

[23] L. Kurgan, L. Homaeian, Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy, Pattern Recognition 39 (12) (2006) 2323–2343.

[24] L. Kurgan, M. Rahbari, L. Homaeian, Impact of the predicted protein structural content on prediction of structural classes for the twilight zone proteins, 5th International Conference on Machine Learning and Applications, (2006) 180–186.

[25] K.D. Kedarisetti, L. Kurgan, S. Dick, Classifier ensembles for protein structural class prediction with varying homology, Bio-

chemical and Biophysical Research Communications 348 (2006) 981–988.
[26] K.D. Kedarisetti, L. Kurgan, S. Dick, A comment on prediction of protein structural classes by a new measure of information discrepancy', Computational Biology and Chemistry 30 (5) (2006) 393–394.
[27] J. Kyte, R.F. Doolitle, A simple method for displaying the hydropathic character of a protein, Journal of Molecular Biology 157 (1982) 105–132.
[28] N. Landwehr, M. Hall, E. Frank, Logistic model trees, Machine Learning Journal 59 (1–2) (2005) 161–205.
[29] X. Li, X. Pan, New method for accurate prediction of solvent accessibility from protein sequence, Proteins 42 (1) (2001) 1–5.
[30] Z. Lin, X. Pan, Accurate prediction of protein secondary structural content, Journal of Protein Chemistry 20 (3) (2001) 217–220.
[31] K. Lin, V.A. Simossis, W.R. Taylor, J. Heringa, A simple and fast secondary structure prediction method using hidden neural networks, Bioinformatics 21 (2) (2005) 152–159.
[32] A. Murzin, S. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of protein database for the investigation of sequence and structures, Journal of Molecular Biology 247 (1995) 536–540.
[33] S.M. Muskal, S.-H. Kim, Predicting protein secondary structure content: a tandem neural network approach, Journal of Molecular Biology 225 (1992) 713–727.
[34] G. Pollastri, A. McLysaght, Porter: a new, accurate server for protein secondary structure prediction, Bioinformatics 21 (8) (2005) 1719–1720.
[35] B. Rost, Twilight zone of protein sequence alignments, Protein Engineering 2 (1999) 85–94.
[36] J. Ruan, K. Wang, J. Yang, L. Kurgan, K. Cios, Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences, Artificial Intelligence in Medicine 35 (1–2) (2005) 9–35.
[37] C. Sander, R. Schneider, Database of homology-derived structures and the structural meaning of sequence alignment, Proteins 9 (1991) 56–68.
[38] H.B. Shen, J. Yang, X.-J. Liu, K.C. Chou, Using supervised fuzzy clustering to predict protein structural classes, Biochemical and Biophysical Research and Communications 334 (2005) 577–581.
[39] Z.-X. Wang, Z. Yuan, How good is the prediction of protein structural class by the component-coupled method? Proteins 38 (2000) 165–175.
[40] I. Witten, E. Frank, Data mining: practical machine learning tools and techniques, Second ed., Morgan Kaufman, San Francisco, 2005.
[41] X. Xiao, S. Shao, Z. Huang, K.C. Chou, Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor, Journal of Computational Chemistry 27 (4) (2006) 478–482.
[42] X. Yang, B. Wang, Weave amino acid sequences for protein secondary structure prediction, 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (2003) 80–87.
[43] C.T. Zhang, Z. Zhang, Z. He, Prediction of the secondary structure contents of globular proteins based on three structural classes, Journal of Protein Chemistry 17 (1998) 261–272.
[44] C.T. Zhang, Z.S. Lin, Z. Zhang, M. Yan, Prediction of helix/strand content of globular proteins based on their primary sequences, Protein Engineering 11 (11) (1998) 971–979.
[45] Z.D. Zhang, Z.R. Sun, C.T. Zhang, A new approach to predict the helix/strand content of globular proteins, Journal of Theoretical Biology 208 (2001) 65–78.
[46] G.P. Zhou, An intriguing controversy over protein structural class prediction, Journal of Protein Chemistry 17 (8) (1998) 729–738.