WILEY InterScience®
DISCOVER SOMETHING GREAT

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Prediction of protein secondary structure content for the twilight zone sequences

Leila Homaeian,[1] Lukasz A. Kurgan,[1]* Jishou Ruan,[2] Krzysztof J. Cios,[3] and Ke Chen[1]

[1] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada
[2] Chern Institute of Mathematics, College of Mathematical Science and LPMC, Nankai University, Tianjin, People's Republic of China
[3] Department of Computer Science and Engineering, University of Colorado at Denver and Health Sciences Center, Denver, Colorado

ABSTRACT

Secondary protein structure carries information about local structural arrangements, which include three major conformations: α-helices, β-strands, and coils. Significant majority of successful methods for prediction of the secondary structure is based on multiple sequence alignment. However, multiple alignment fails to provide accurate results when a sequence comes from the twilight zone, that is, it is characterized by low (<30%) homology. To this end, we propose a novel method for prediction of secondary structure content through comprehensive sequence representation, called PSSC-core. The method uses a multiple linear regression model and introduces a comprehensive feature-based sequence representation to predict amount of helices and strands for sequences from the twilight zone. The PSSC-core method was tested and compared with two other state-of-the-art prediction methods on a set of 2187 twilight zone sequences. The results indicate that our method provides better predictions for both helix and strand content. The PSSC-core is shown to provide statistically significantly better results when compared with the competing methods, reducing the prediction error by 5–7% for helix and 7–9% for strand content predictions. The proposed feature-based sequence representation uses a comprehensive set of physicochemical properties that are custom-designed for each of the helix and strand content predictions. It includes composition and composition moment vectors, frequency of tetra-peptides associated with helical and strand conformations, various property-based groups like exchange groups, chemical groups of the side chains and hydrophobic group, auto-correlations based on hydrophobicity, side-chain masses, hydropathy, and conformational patterns for β-sheets. The PSSC-core method provides an alternative for predicting the secondary structure content that can be used to validate and constrain results of other structure prediction methods. At the same time, it also provides useful insight into design of successful protein sequence representations that can be used in developing new methods related to prediction of different aspects of the secondary protein structure.

Proteins 2007; 69:486–498.

Key words: protein structure; secondary protein structure; secondary structure content; twilight zone; low sequence homology.

INTRODUCTION

Prediction of the secondary structure content, defined as the percentage amount of helices and strands in a protein, provides useful information for characterization of the overall protein structure. The dictionary of secondary structures of proteins (DSSP) annotates each amino acid (AA) as belonging to one of eight secondary structure types[1]: H (helix), G (3_{10}-helix), I (pi-helix), B (residue in isolated-bridge), E (extended strand), T (hydrogen bond turn), S (bend), and "_" (any other structure). Typically these eight secondary structure types are reduced to just three groups[2]: helix (which includes types H, G, and I), strand (which includes types E and B), and coil (which includes T, S, and the others). Although the protein secondary structure content prediction can be performed for eight-state representation,[3–6] majority of prior attempts, including this work, address the three-state problem.

The first secondary content prediction effort was undertaken in early 1970s when a multiple linear regression (MLR) model was used to predict the content utilizing the composition vector-based sequence representation for a small set of 18 proteins.[7] It was not until 1990s when another content prediction approach was proposed.[8] The authors used composition vector, molecular weight of the sequence and absence/presence of bound Heme group to represent protein sequences and two neural networks to perform the prediction. Still another method used the composition vector representation and analytic vector decomposition technique to predict the content.[9] In late 1990s a MLR model was used on the sequence representation that for the first time used auto-correlation functions based on hydrophobicity.[10] Similar methods, which

Grant sponsors: NSERC, MITACS, Liuhui Center for Applied Mathematics, NSFC, Butcher Foundation.

*Correspondence to: Lukasz A. Kurgan, Department of Electrical and Computer Engineering, 2nd floor, ECERF (9107 116 Street), University of Alberta, Edmonton, AB, Canada T6G 2V4. E-mail: lkurgan@ece.ualberta.ca

Received 13 November 2006; Revised 15 February 2007; Accepted 22 March 2007
Published online 10 July 2007 in Wiley InterScience (www.interscience.wiley. com).
DOI: 10.1002/prot.21527

Table I

Comparison of Example Predicted Content Values Based on the Predicted Secondary Structure with PSI-PRED and YASPIN Methods and the Results of the Proposed PSSC-Core Method

	Three-state accuracy (Q_3) of predicted secondary structure		Secondary structure content							
			True content		Predicted based on PSI-PRED		Predicted based on YASPIN		Predicted by PSSC-Core	
PDB ID	PSI-PRED	YASPIN	Helix	Strand	Helix	Strand	Helix	Strand	Helix	Strand
1UFXA	88.3	82.5	0.175	0.282	0.146	0.33	0.155	0.34	0.166	0.283
1K83K	85.8	81.7	0.367	0.217	0.358	0.267	0.333	0.275	0.356	0.216
1GRCA	82.1	77.8	0.311	0.231	0.344	0.274	0.335	0.278	0.312	0.237
1GLLO	79.6	74.5	0.343	0.228	0.385	0.156	0.395	0.226	0.348	0.224
1HZTA	77.4	67.4	0.226	0.253	0.258	0.263	0.316	0.284	0.232	0.245

additionally introduced auto-correlation functions to represent protein sequences, and also used MLR models were developed in early 2000s.[11,12] Pilizota et al. performed feature selection on the composition vector and used the resulting representation and the MLR models to perform prediction.[13] Recently, a neural network approach proposed a novel composition moment vector-based sequence representation.[14] Several researchers investigated impact of a priori knowledge of structural classes on the quality of the content prediction. These methods generated separate MLR models for each structural class.[11,15,16] Their main drawback is that they require knowledge of the structural class of the input sequence to perform prediction. This could be either inferred based on the known secondary structure, or predicted, but structural class prediction is difficult and is characterized by relatively low accuracy.[17,18]

While majority of the protein structure prediction methods use multiple sequence alignment, the secondary structure content prediction is usually performed based on classification of the sequences that are converted into a feature-based representation.[11,12] Low homology sequences pose a substantial challenge for structure prediction, since sequence alignment requires at least ∼30% homology between the query protein and protein(s) used to correctly predict its structure.[19] The proteins characterized by a lower 20–30% homology with sequences that are used to predict their structure are called twilight zone proteins.[20] More than 95% of all sequence pairs detected in the twilight zone have different structures,[20] which significantly reduces the quality of prediction. For instance, prediction of the secondary structure for homologous sequences by the state-of-the-art alignment-based methods yields about 80% accuracy,[21] while for the twilight zone sequences it drops to only 65–68%.[22] 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.[17] Table I shows example results for several twilight zone proteins used in this paper. The results include accuracy of the predicted secondary structure with the best performing PSI-PRED method and recently developed (using twilight zone proteins) YASPIN methods,[22] the content values inferred from the predicted secondary structures, and the predictions of the proposed method. We stress that although the secondary structure prediction methods and the content prediction methods have different goals and cannot be directly compared (the secondary structure prediction methods use proteins that we predict to build their prediction model), the content information could be useful in improving the prediction of the secondary structure for the low homology proteins.

This paper aims to improve accuracy of content prediction for the twilight zone proteins without using the sequence alignment. The prediction is performed in two steps. First, the protein sequences of various lengths are converted into a fixed size feature vector. Second, the feature vectors are fed into a prediction model to obtain the content values. Similarly to the other prior works, we use the MLR model to perform prediction. However, while prior prediction methods use a simple composition vector and a few feature sets as the sequence representation, our method proposes new, comprehensive, custom-designed sequence representation that improves the quality of prediction. The PSSC-core method combines (1) a set of features obtained via feature selection from a comprehensive set of features that were used in prior methods for prediction of secondary structure content, structure, structural class and function, and (2) a set of newly proposed features.

The method uses two MLR models, one for helix content prediction that uses 46 features and another for strand content prediction that uses 88 features to represent the sequences.

MATERIALS AND METHODS

Datasets

As the proposed prediction method aims at predicting secondary structure content for the twilight zone proteins the corresponding datasets are characterized by controlled levels of sequence homology.

The method was tested on the PDBSelect25 dataset, which is a nonredundant representative subset of the protein data bank (PDB) characterized by 25% average sequence homology.[23] This dataset also excludes PDB sequences that satisfy the following criteria: (1) contain more than 5% of nonstandard AAs, that is residues other than the 20 common AAs that are typically denoted as UNK in PDB files; (2) have less than 30 AAs; and (3) are measured with a resolution greater than 3.5 Å and with R-factor greater than 30%. We used PDBSelect25 version from November 2004 that includes 2485 sequences. This dataset was further processed to exclude sequences with any number of nonstandard AAs. We note that none of the remaining sequences includes helix fragments shorter than three residues. As a result, the final dataset includes 2187 sequences.

The recently proposed prediction method[12] used a dataset of 704 sequences characterized by 30% homology.[24] Since this dataset is relatively small and includes sequences with higher homology, it was combined with the PDBSelect25 to show impact of the increased homology on the prediction quality. Authors of Ref. 12 provided us with a list of 681 sequences from the original dataset. This dataset was preprocessed using the same criteria as for the PDBSelect25 dataset resulting in removal of 39 sequences. After combining these sequences with the PDBSelect25 dataset, the CD-HIT program[25] was used with the lowest homology filtration setting to remove sequences with a homology higher than 40%. The resulting dataset, referred to as LinPanPlus25, includes 2483 sequences.

The PDBSelect90 list[23] was used to create dataset used to optimize parameters of our prediction method. The original list from November 2004 included 8595 highly homologous proteins, which were filtered using the same criteria as for the PDBSelect25 dataset, resulting in 7544 sequences. Sequences with a homology level greater than 40% to sequences in PDBSelect25 and the dataset of 704 sequences used in Ref. 12 were removed using the CD-HIT to eliminate bias between the design (performed using this dataset) and testing (performed using PDBSelect25 and LinPanPlus25 datasets) of the proposed method. The resulting dataset, referred to as *PDBSelect90LH*, includes 3987 sequences. This dataset was further randomly divided into two subsets: 80% of the original dataset was used to create dataset D1 and the remaining 20% to create dataset D2.

Prediction model

The PSSC-core method performs prediction in two steps. First, the protein sequences are represented as feature vectors. Next, two multiple linear regressions models are used to predict helix and strand contents, see Figure 1.

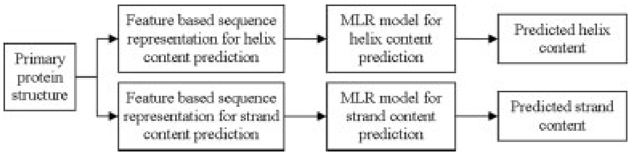

Primary protein structure
Feature based sequence representation for helix content prediction
Feature based sequence representation for strand content prediction
MLR model for helix content prediction
MLR model for strand content prediction
Predicted helix content
Predicted strand content

Figure 1
The overall structure of the PSSC-core method.

We note that nonlinear regression models, which includes quadratic, cubic, exponential, and Fourier transform coefficients were also tested and shown to provide lower prediction quality than the linear models.[26]

The MLR models for helix and strand content prediction are defined as:

$$y_\alpha = a_0 + \sum_{i=1}^{n_\alpha} a_i f_i$$

$$y_\beta = b_0 + \sum_{i=1}^{n_\beta} b_i f_i$$

where y_α and y_β denote the predicted α-helix and β-strand content, f denotes protein feature vector (also referred to as the regression predictors) that includes n_α and n_β dimensions for helix and strand content predictions, respectively, and $a_0,a_1\ldots a_n$ (and $b_0,b_1, \ldots ,b_n$) are regression coefficients estimated using training data. Since helix and strand content values are real numbers normalized to the interval between zero and one, any negative number prediction is rounded to zero, while predictions greater than one are rounded to one.

Evaluation of the model

The models were evaluated using resubstitution and cross-validation. During resubstitution (also referred to as the self consistency test) both training and testing data are the same and therefore the results tend to overestimate the quality of the model. The corresponding results are reported to maintain consistency with the prior research, although they should not be used for evaluative or comparative purposes. To ensure statistical validity of the results we performed 10-fold cross validation. In this cross validation, out of the 10 equal size subsets of the data, one is retained to test the prediction model, and the remaining nine are used to generate the model. The cross validation process is repeated 10 times, with each of the 10 subsets used exactly once as the test data. The results from these 10-folds are averaged to produce a robust estimate of the quality of the model.

DOI 10.1002/prot

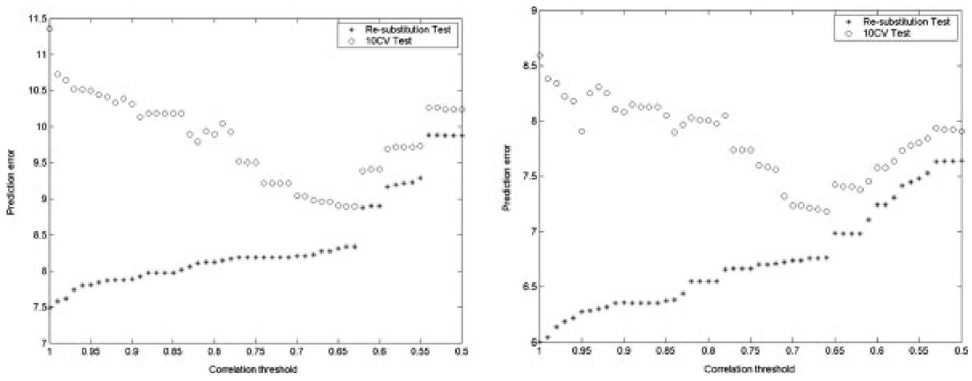


Figure 2

Prediction quality in function of correlation threshold for dataset D2; graph on the left (right) corresponds to helix (strand) content prediction.

Generation of the model

The models are generated using training data using these steps: (1) sequences from the training data are represented by feature vectors (46 and 88 features for helix and strand content prediction, respectively) which include the true content values; (2) pairwise collinearity is computed for each pair of attributes, and for the collinear pairs the feature with a higher correlation with the target content is kept, while the other feature is removed; (3) best-fitting, with respect of the remaining features, of the coefficients of the linear model is performed.

To be well-defined the MLR model requires that the input features are not correlated. To identify correlated features the MLR model was tested using 10-fold cross validation on dataset D2 to establish cut-off thresholds for the correlation coefficient values. The correlation coefficient threshold was changed from 1.00 to 0.50 for helix and strand separately, and the threshold leading to the least 10-fold cross validation error was picked for each secondary structure. Figure 2 shows how the resubstitution and 10-fold cross validation error change as the correlation threshold decreases. Thresholds 0.64 and 0.66 were selected for helix and strand, respectively.

Feature selection

The first step of the prediction procedure requires feature-based sequence representation. The proposed representation was developed based on synergic combination of a selected subset of features that were used in past studies related to protein structure analysis and prediction, and a set of newly proposed features. The subset of features is selected based on a feature selection method described below.

The used forward feature selection method first selects the feature that has the highest absolute correlation coefficient value with the predicted content values. The second feature is added to the model such that it leads to the best two-variable model given that the first feature is already included in the model.[27] The method proceeds by adding one feature at a time provided that the feature leads to the least residual sum of squares (RSS) compared with other features, which are currently not in the model:

$$\mathrm{RSS} = \sum_{i=1}^{s} (y_i^{\mathrm{obs}} - y_i^{\mathrm{pred}})^2$$

where y_i^{obs} and y_i^{pred} are the observed (true) and predicted secondary structure contents for sequence i, respectively and S is the total number of sequences in the training dataset. Assuming that k features are already in the model; feature j is added to the model if the following inequality, which is based on the F-statistic, is satisfied[27]

$$\mathrm{FStat}_j = \max\left(\frac{\mathrm{RSS}_k - \mathrm{RSS}_{k+1}}{\mathrm{RMS}_{k+1}}\right) > \mathrm{FStat}_{\mathrm{in}}$$

where RSS_k is the residual sum of squares for the current model, RSS_{k+1} and RMS_{k+1} are the residual sum of squares and residual mean of squares, respectively, given that feature j is included in the model, and $\mathrm{FStat}_{\mathrm{in}} = 2$ is used.[27]

Protein sequence representation

The prediction of secondary protein content is usually performed with an intermediate step, in which the primary sequence is converted into its feature representation. The existing protein secondary structure content prediction methods use a limited set of features to describe the primary sequence,[7,8,10–14,16] while other methods, such as those for prediction of protein structure, structural class or function use a more diverse and larger number of features.[17,18,28–35] As a follow up of our previous feature selection based approach to protein

Table II
Features Used to Encode Protein Sequences and their Applications

Features	Abbr.	Application(s)	Reference(s)
Index-based			
Sequence length	N	Protein content, structural class and function prediction	8, 17, 18, 28, 31
Average molecular weight (Table IV)	MolW		
Average isoelectric point (Table IV)	pI		
Auto-correlation functions based on FH$_i$, EH$_i$, and M_i indices (Table IV)	A_n^{FH}, A_n^{EH}, A_n^{M}	Protein content and structural class prediction; characterization of degree of similarity between three-dimensional protein structures	10–12, 17, 18, 37, 38
Average hydrophobicities based on FH$_{i,}$ and EH$_i$ indices	$H_{\mathrm{avr}}^{\mathrm{EH}}$, $H_{\mathrm{avr}}^{\mathrm{FH}}$	Protein structural class prediction	18
Sum of hydrophobicities based on FH$_{i,}$ and EH$_i$ indices	$H_{\mathrm{sum}}^{\mathrm{FH}}$, $H_{\mathrm{sum}}^{\mathrm{EH}}$	Protein structural class prediction	18
Sum of three-running average of hydrophobicities of FH$_{i,}$ and EH$_i$ indices	$H_{\mathrm{sum3}}^{\mathrm{FH}}$, $H_{\mathrm{sum3}}^{\mathrm{EH}}$	Protein structural class prediction	18
Composition moment vector			
Composition vector	CV	Protein structure, structural class, and content prediction	7–11, 14–18, 34, 39, 40
First and second order composition moment vector	CMV^1, CMV^2	Protein content and structural class prediction	14, 17, 18
Property groups			
Hydrophobicity groups	HG	Protein function prediction; characterization of structural and functional relationships	18, 28, 31
R groups	RG	Protein structural class and content prediction	18, 32
Exchange groups	XG	Protein family and structural class prediction	18, 29, 32
Electronic groups	EG	Protein structure and structural class prediction	17, 18, 33
Chemical groups	CG		
Other groups	OG	Protein function and structural class prediction; characterization of structural, and functional relationships	18, 28, 31

content prediction[36] this paper performs a comprehensive feature selection from the aggregated set of physicochemical features. Most importantly, a new set of features is proposed and added to the selected features.

Feature selection from the existing feature sets

The feature sets proposed in prior research and their original applications are summarized in Table II.

The features can be divided into three groups:

Index-based feature sets. These feature sets are derived based on physicochemical AA indices, see Table III, and include molecular weight, average isoelectric point, and auto-correlation functions based on hydrophobicity and side chain masses. The molecular weight, MolW, of a protein sequence is the result of adding up the molecular weight MolW$_i$ (residue average) values of its residues plus the mass of a water molecule (MolW$_{\mathrm{H_20}}$) that is approximately 18 Da.

$$\mathrm{MolW} = \mathrm{MolW}_{\mathrm{H_20}} + \sum_{i=1}^{N} \mathrm{MolW}_i$$

where N is the length of the protein, that is number of AAs.

The average isoelectric point, pI, of a protein sequence is computed based on the average isoelectric point pI_i values of its residues. The pI value shows the pH at which a molecule carries no net electric charge and thus it is immobile in an electric field[41]:

$$\mathrm{p}I = \frac{1}{N} \sum_{i=1}^{N} \mathrm{p}I_i$$

An order n auto-correlation function is computed by summing up the products of AA indices a_i of every pair of residues separated by n residues.

$$A_n^a = \frac{1}{N-n} \sum_{i=1}^{N-n} a_i a_{i+n}$$

The autocorrelations are computed for:

- Two hydrophobicity indices: the Fauchere–Pliska's (FH) index[42] and the Eisenberg's (EH) index.[43]
- The relative side-chain masses M index.[12]
- Hydropathy index, Hp, proposed in Ref. 44, where it was used to identify the hydrophilic and hydrophobic regions of a protein.

Following on previous research we use six auto-correlation functions based on FH, EH, and M,[12] that is $n = [1\ldots6]$, and nine function based on Hp.

DOI 10.1002/prot

Table III
List of Physicochemical Amino Acid Indices Used to Derive the Protein Sequence Representation

			Physicochemical index						
Amino acid	Code	Index	*MolW*	*pI*	*FH*	*EH*	*M*	*Hp*	*E*
Alanine	A	1	71.0791	6.01	0.42	0.62	0.115	1.8	4.8
Cysteine	C	2	103.1437	5.07	1.34	0.29	0.36	2.5	0.6
Aspartate	D	3	115.0887	2.77	−1.05	−0.9	0.446	−3.5	0.5
Glutamate	E	4	129.1157	3.22	−0.87	−0.74	0.55	−3.5	1.1
Phenylalanine	F	5	147.1772	5.48	2.44	1.19	0.7	2.8	5.3
Glycine	G	6	57.0521	5.97	0	0.48	0.00076	−0.4	3.0
Histidine	H	7	137.1414	7.59	0.18	−0.4	0.63	−3.2	0.6
Isoleucine	I	8	113.16	6.02	2.46	1.38	0.13	4.5	17.8
Lysine	K	9	128.1792	9.74	−1.35	−1.5	0.48	−3.9	1.67
Leucine	L	10	113.16	5.98	2.32	1.06	0.13	3.8	12.16
Methionine	M	11	131.1977	5.47	1.68	0.64	0.577	1.9	0.67
Asparagine	N	12	114.104	5.41	−0.82	−0.78	0.446	−3.5	0.46
Proline	P	13	97.1171	6.48	0.98	0.12	0.323	−1.6	−0.08
Glutamine	Q	14	128.131	5.65	−0.3	−0.85	0.55	−3.5	0.54
Arginine	R	15	156.188	10.76	−1.37	−2.53	0.777	−4.5	1.07
Serine	S	16	87.0784	5.68	−0.05	−0.18	0.238	−0.8	1.9
Threonine	T	17	101.1054	5.87	0.35	−0.05	0.346	−0.7	6.6
Valine	V	18	99.133	5.97	1.66	1.08	0.33	4.2	35.25
Tryptophan	W	19	186.2139	5.89	3.07	0.81	1	−0.9	0.65
Tyrosine	Y	20	163.1756	5.67	1.31	0.26	0.82	−1.3	5.27

The sum, average and three-point running average of hydrophobicity indices was also computed[18]:

$$H^{b}_{\mathrm{sum}} = \sum_{i=1}^{N} b_i$$

$$H^{b}_{\mathrm{avr}} = \frac{\sum\limits_{i=1}^{N} b_i}{N}$$

$$H^{b}_{\mathrm{sum3}} = \sum_{i=1}^{N-3} \Big(\sum_{j=1}^{i+3} b_j\Big)/3$$

where $b_i = \{\mathrm{FH}, \mathrm{EH}\}$.

Finally, the cumulative density functions based on Feuchere–Pliska's and Eisenberg's hydrophobic indices were computed as:

$$H\mathrm{Cum}^{b}_{n} = \frac{\sum_{i=1}^{N-n}\left(\sum_{j=1}^{i} b_j\right)\left(\sum_{j=1}^{i} b_j\right)}{N-n}$$

where b_j is the AA index value for jth AA, N is the length of the sequence, and $n = [1\ldots 6]$. The functions were computed based on the two hydrophobic indices, that is, $b_j = \{\mathrm{FH}, \mathrm{EH}\}$ and resulted in total of 12 features.

Composition vector and composition moment vector. Composition vector (CV) is defined as the composition percentage of each residue in the primary sequence. Unlike composition vector, composition moment vector (CMV) takes into account the position of each residue in the sequence:

$$\mathrm{CMV}^{k}_{i} = \frac{\sum_{j=1}^{n_i} n^{k}_{ij}}{\prod_{d=0}^{k} (N-d)}$$

n_{ij} represents the jth position of the ith AA, n_i is the frequency of ith AA in the sequence, and k is the order of the CMV. We apply CMVs for $k = 0, 1, 2$. Note that CMV for $k = 0$ reduces to CV. The composition vector was used extensively for both protein structure and content prediction (Table III), while CMV was recently proposed for the protein content and structural class prediction.[14,17,18]

Property groups. AAs can be clustered based on their common properties, see Table IV, and composition is computed for each of the groups and subgroups. Hydrophobicity group includes hydrophilic AAs, which are water-soluble with ionized or polar side chains. Usually they are located at the surface of a water-soluble protein. In contrast, hydrophobic AAs are slightly soluble or insoluble. R group classification is based on molecular weight, hydropathy, and isoelectric point.[32] Exchange groups are supported by statistical studies and cluster AAs based on accepted point mutation to represent conservative replacements through revolution. Electronic group classification is based on tendency of AAs to accept or donate electrons.[33] Chemical groups are defined based on composition of chemical group that constitute the side chains,[33] see Table V. Finally, other

Table IV
Property Groups of Amino Acids

Groups	Subgroups	AAs	Groups	Subgroups	AAs
R groups	Nonpolar aliphatic	AVLIMG	Hydrophobicity groups	Hydrophobic	VLIMAFPWYCG
	Polar uncharged	SPTCNQ		Hydrophilic basic	KHR
	Positively charged	KHR		Hydrophilic acidic	DE
	Negative	DE		Hydrophilic polar with uncharged side chain	STNQ
	Aromatic	FYW			
Exchange groups	(A)	C	Electronic groups	Electron donor	DEPA
	(C)	AGPST		Weak electron donor	VLI
	(D)	DENQ		Electron acceptor	KNR
	(E)	KHR		Weak electron acceptor	FYMTQ
	(F)	ILMV		Neutral	GHWS
	(G)	FYW		Special AA	C
Other groups	Charged	DEKHRVLI	Other groups	Tiny	AG
	Polar	DEKHRNTQSYW		Bulky	FHWYR
	Aromatic	FHWY		Polar-uncharged	NQ

groups are defined based on molecular weights, that is tiny, small, and bulky AAs are less than 80 Da, between 80 and 101 Da, and more than 120 Da, respectively,[28] polarity, aromaticity, and charge.

The forward feature selection was performed on the above features using the PDBSelect90LH dataset separately for helix and strand content. The following 44 features were selected to represent sequences for helix content prediction:

- average isoelectric point of the sequence;
- composition vector for Alanine (A), Phenylalanine (F), and Leucine (L);
- first order composition moment vector for Alanine (A), Isoleucine (I), Leucine (L), Methionine (M), Threonine (T), and Tryptophan (W);
- second order composition vector for Alanine (A), Glutamate (E), Methionine (M), and Threonine (T);
- C exchange group;
- following four chemical groups of the side chains: CAROM, CH_2, CO, and OH.
- small and polar-uncharged other groups;
- auto-correlation based on Feuchere-Pliska's hydrophobic index for $n = 1, 2, 3, 4, 5$, and 6;
- sum over the Feuchere-Pliska's hydrophobic index;
- auto-correlation based on Eisenberg's hydrophobic index for $n = 1, 2, 3, 4$, and 6;
- auto-correlation based on average side chain masses index for $n = 1, 2$, and 6;
- cumulative density for Eisenberg's hydrophobic index for $n = 1, 4$, and 5;
- auto-correlation based on hydropathy index for $n = 2$, 3, 4, 7, and 8.

Similarly, the following 67 features were selected to represent sequences for strand content prediction:

- composition vector for Alanine (A), Cysteine (C), Isoleucine (I), Methionine (M), Threonine (T), and Valine (V);
- first-order composition vector for Alanine (A), Cysteine (C), Glutamate (E), Glycine (G), Histidine (H),

Table V
Chemical Composition of the Side Chains

AA	Associated chemical groups	AA	Associated chemical groups
A	CH CO NH CH_3	M	CH CO NH CH_2 CH_2 S CH_3
C	CH CO NH CH_2 SH	N	CH CO NH CH_2 CO C NH_2
D	CH CO NH CH_2 CO COO^-	P	CHRING CO NHRING CH_2RING CH_2RING CH_2RING
E	CH CO NH CH_2 CH_2 CO COO^-	Q	CH CO NH CH_2 CH_2 CO C NH_2
F	CH CO NH CH_2 CAROM CHAROM CHAROM CHAROM CHAROM CHAROM	R	CH CO NH CH_2 CH_2 CH_2 NH C NH_2 NH_2^+
G	CH_2 CO NH	S	CH CO NH CH_2 OH
H	CH CO NH CH_2 CAROM CHAROM N CHAROM NH	T	CH CO NH CH CH_3 OH
I	CH CO NH CH_2 CH CH_3 CH_3	V	CH CO NH CH CH_3 CH_3
K	CH CO NH CH_2 CH_2 CH_2 CH_2 NH_3^+	W	CH CO NH CH_2 CAROM CAROM CAROM NH CHAROM CHAROM CHAROM CHAROM CHAROM
L	CH CO NH CH_2 CH CH_3 CH_3	Y	CH CO NH CH_2 CAROM CHAROM CHAROM CHAROM CHAROM CAROM OH

DOI 10.1002/prot

Isoleucine (I), Leucine (L), Methionine (M), Asparagine (N), Glutamine (Q), and Valine (V);
- second-order composition vector for Alanine (A), Cysteine (C), Aspartate (D), Glutamate (E), Glycine (G), Histidine (H), Isoleucine (I), Lysine (K), Leucine (L), Methionine (M), Asparagine (N), Glutamine (Q), Arginine (R), and Valine (V);
- negative R-group;
- hydrophobic hydrophobicity group;
- following two chemical groups of the side chains: CH_2 and OH;
- polar, tiny, bulky, and polar-uncharged other groups;
- auto-correlation based on Feuchere–Pliska's hydrophobic index for $n = 1, 2, 3, 4$, and 6;
- average over the Feuchere–Pliska's hydrophobic index;
- sum over three-point averages of the Feuchere–Pliska's hydrophobic index;
- auto-correlation based on Eisenberg's hydrophobic index for $n = 1, 2, 3, 4, 5$, and 6;
- auto-correlation based on average side chain masses index for $n = 1, 2, 4$, and 6;
- cumulative density for Feuchere–Pliska's hydrophobic index for $n = 1, 3, 4$, and 6;
- auto-correlation based on hydropathy index for $n = 2, 3, 4, 5, 7, 8$, and 9.

The selected features confirm high quality of the composition and composition moment vectors (13 and 31 features based on these feature sets were selected for prediction of helix and strand content, respectively), although different and relatively small subsets of AAs for these features were selected for each of the predictions (secondary structures). The high value of hydrophobicity based features was also confirmed (20 and 25 features computed based on hydrophobicity were selected for prediction of helix and strand content, respectively). Both hydrophobic indices used in prior research provided useful features for prediction of both helix and strand contents.[11,12] Finally, information related to the side chains was also found useful. The selected features in both cases include information about the CH_2 and OH chemical groups of the side chains, while additionally composition of CAROM and CO groups was selected in case of helices.

New feature sets

The feature based sequence representation used by the PSSC-core uses two more features for the helix prediction representation and 21 more for the strand prediction representation. They were based on the following two feature sets:

- auto-correlation functions based on a new index developed based on statistical analysis of conformational patterns of β-sheets;
- frequency of tetra-peptides associated with helical and strand conformations.

Both feature sets were applied for prediction of the strand content, while the second one was applied only for the helix content prediction. These features were not processed by feature selection since they were specifically designed and tested to improve the content prediction.

Although the hydrophobicity, hydropathy, and side chain masses based autocorrelation functions reflect local, with respect to the sequence, structural arrangements, the long range interactions that are characteristic for β-sheets are not covered by the feature sets proposed in the past. Therefore, a new index was developed based on statistical analysis of long range interactions between strands that form a β-sheet. The index was computed based on complementary pairs of AAs A_i and A_j that belong to two β-strands that are connected by hydrogen bonds to form the sheet. The conditional probability distributions of these pairs are defined as:

$$p(A_i : A_j) = \frac{N(A_i : A_j)}{N}$$

where $N(A_i : A_j)$ denotes the number of pairs $A_i : A_j$ in β-sheets collected based on PDB release No. 103, and N is the total number of these pairs.

The index values, E_i, are computed as:

$$E_i = \frac{100p(A_1)L(A_i)}{\sum_{i=1}^{20} p(A_1)L(A_i)}$$

where $L(A_i) = \sum_{j=1}^{20} p(A_i : A_j) \log_2 \frac{p(A_i:A_j)}{p(A_i)p(A_j)}$, and $p(A_i)$ and $p(A_j)$ are the probabilities of A_i and A_j in β-sheets collected based on PDB release No. 103. The corresponding index values, denoted as E, are shown in Table III. The index was used to compute auto-correlation functions A_n^E for $n = [1,\ldots,19]$ and thus resulted in adding 19 new features for the sequence representation used for prediction of strand content. The optimal number of autocorrelation functions was selected based on maximization of the prediction quality for the PDBSelect25 dataset (using 10-fold cross validation).

The final set of features that are based on polypeptide composition was derived based on statistical analysis of the D1 and D2 datasets. The objective was to find a set of polypeptides of the same length that are frequently observed in each of the helix and strand structures. For each secondary structure one feature was computed as the sum of frequencies of all frequent polypeptides in the corresponding protein sequence. We experimented with di-, tri-, and tetra-peptides, and concluded that the best results are achieved for the tetra-peptides.[26]

The procedure to extract the set of tetra-peptides is as follows:

- scan dataset D1 and report the frequency of each polypeptide observed in helix (strand) across the entire dataset;
- normalize each frequency based on the frequency of the same polypeptide when disregarding the secondary structure;
- sort polypeptides with regard to the normalized frequencies;
- based upon a threshold, keep the most frequent polypeptides.

The cut-off thresholds were found based on 10-fold cross validation with dataset D2. For each threshold, a corresponding set of tetra-peptides was generated from D1 and the corresponding two attributes together with the remaining attributes that constitute the proposed feature based sequence representation were used to perform 10-fold cross validation tests with dataset D2. The thresholds for helix content prediction are as follows: $\mu + 0.7\sigma'$ for helix and $\mu + 0.9\sigma'$ for strand. The thresholds for strand content prediction are as follows: $\mu + 0.5\sigma'$ for helix and $\mu + 0.3\sigma'$ for strand; where μ and σ' are the mean and standard deviation of the frequencies, respectively.

The final feature-based sequences representation, which incorporates features from feature selection and the new feature sets, includes 46 (44 from the feature selection and 2 from the new feature sets) features for helix content prediction model and 88 (67 from the feature selection and 21 from the new feature sets) features for strand content prediction model.

RESULTS AND DISCUSSION

Experimental setup

We compared our PSSC-core with methods proposed by Lin and Pan[12] and Zhang et al.[11]; these papers provide a solid baseline for comparison as they describe recent and the best performing methods. They were comprehensively tested on a relatively large protein set (over 700 sequences) and compared with many methods including neural networks,[8] MLR models[7,11] and vector decomposition.[9] The Lin and Pan's method was shown to provide the best predictions.[12] Analysis of test results that accompany the most recent neural network and MLR-based prediction methods[13,14] indicates that they are characterized by lower prediction quality than the Lin–Pan's method. Additionally, we also compare our method with the most accurate method, which was proposed by Zhang et al.,[11] that uses a priori structural class information assuming, and similarly to[12] we assume that the a priori knowledge of structural class is not provided.

Following[11,12] the mean absolute error is used to measure the accuracy of a content prediction model separately for helix and strand:

$$e = \frac{\sum_{i=1}^{S} \left| y_i^{\mathrm{obs}} - y_i^{\mathrm{pred}} \right|}{S}$$

where S is the number of sequences in the dataset for which the prediction is performed, y_i^{obs} and y_i^{pred} are the observed (true) and predicted secondary structure contents for sequence i, respectively. The standard deviation of the prediction error, σ, is also reported:

$$\sigma = \sqrt{\sum_{i=1}^{S} \frac{(e - |y_i^{\mathrm{obs}} - y_i^{\mathrm{pred}}|)^2}{S-1}}$$

On the basis of the data used during the tests, we observed that the actual (true) content values range between 0 and 0.93 and between 0 and 0.75 for strands and helices, respectively. Therefore, a given absolute error value for helix content prediction should be considered as relatively larger when compared with the same absolute error for strand content prediction. To this end, we also computed and reported normalized prediction error for both helices and strands

$$p = \frac{\sum_{i=1}^{S} \left| \frac{y_i^{\mathrm{obs}} - y_i^{\mathrm{pred}}}{y_{\max}^{\mathrm{obs}} - y_{\min}^{\mathrm{obs}}} \right| \times 100}{S}$$

where $y_{\max}^{\mathrm{obs}}$ and $y_{\max}^{\mathrm{pred}}$ are the observed (true) maximal and minimal value of the corresponding secondary structure content. This error index quantifies, in percent, the average ratio between the absolute prediction error and the corresponding range of the content values. As a result, it allows to properly scale the errors between the predictions for helices and strands.

Experimental results

The results are reported for two large protein sets, see Table VI. The first dataset (PDBSelect25) includes 2187 twilight zone sequences. The second dataset (LinPanPlus25) that includes 2483 sequences is a result of merging the PDBSelect25 with a smaller dataset of 704 sequences used in Ref. 12. Four methods are compared: (1) method proposed in Ref. 12; (2) method proposed in Ref. 11; (3) MLR method based on feature selection (PSSC-core without newly proposed features); and (4) PSSC-core. Although the resubstitution (in-sample) test results are reported for consistency with prior works, the 10-fold cross validation results are used to compare our method with the method proposed in Refs. 11 and 12.

The PSSC-core method provides very good results for both strand and helix content predictions. Overall, the strand content can be predicted with a higher quality

DOI 10.1002/prot

Table VI

Comparison of Resubstitution and 10-Fold Cross Validation Prediction Results on the PDBSelect25 and LinPanPlus25 Datasets for the Proposed PSSC-Core Method, the Method that Applied only the Feature Selection Based Sequence Representation and the Two Competing Methods

		Helix				Strand			
		Resubstitution		10-fold cross validation		Resubstitution		10-fold cross validation	
Data	Method	$e(\sigma)$	p	$e(\sigma)$	p	$e(\sigma)$	p	$e(\sigma)$	p
PDBSelect25	Lin and Pan's method[12]	0.114 (0.09)	12.3	0.116 (0.09)	12.5	0.086 (0.07)	11.4	0.088 (0.07)	11.7
	Zhang et al.'s method[11]	0.113 (0.09)	12.2	0.115 (0.09)	12.4	0.085 (0.07)	11.4	0.087 (0.07)	11.5
	Feature selection	0.108 (0.09)	11.6	0.111 (0.09)	12.0	0.082 (0.07)	10.9	0.085 (0.07)	11.3
	PSSC-core	0.107 (0.09)	11.5	**0.109 (0.09)**	**11.8**	0.077 (0.06)	10.2	**0.080 (0.07)**	**10.7**
LinPanPlus25	Lin and Pan's method[12]	0.110 (0.09)	11.8	0.112 (0.09)	12.0	0.083 (0.07)	11.1	0.085 (0.07)	11.3
	Zhang et al.'s method[11]	0.109 (0.09)	11.7	0.111 (0.09)	11.9	0.083 (0.07)	11.0	0.084 (0.07)	11.2
	Feature selection	0.104 (0.09)	11.1	0.106 (0.09)	11.4	0.080 (0.07)	10.6	0.082 (0.07)	10.9
	PSSC-core	0.102 (0.09)	11.0	**0.105 (0.09)**	**11.2**	0.075 (0.06)	10.0	**0.078 (0.07)**	**10.4**

than the helix content, even when considering normalized error *p*. This is especially valuable in the context of the secondary structure prediction, in which the helices are usually predicted with higher accuracy than the strands.[22] The results for the PDBSelect25 dataset, which includes twilight zone proteins, are characterized by lower quality than the results for the LinPanPlus25 dataset, which includes some sequences with higher homology. This confirms that content prediction is more difficult for low homology sequences.

Although the differences in errors between the PSSC-core and the two competing methods may seem small (about 1%), they provide relatively large reduction of the overall error values that oscillate around 12%. On the basis of Table VI, for the PDBSelect25 dataset PSSC-core reduces errors by 0.7/12.5 = 6% and by 0.6/12.4 = 5% for helix content prediction when compared with the methods proposed by Lin and Pan and Zhang et al., respectively. For strand content prediction the corresponding improvements equal 1/11.7 = 9% and 0.8/11.5 = 7%. Similarly, for the LinPanPlus25 dataset, the reduction of the error for helix equals 0.8/12 = 7% and 0.7/11.9 = 6%, and for strand it equals 0.9/11.3 = 8% and 0.8/11.2 = 7%, when compared with Lin and Pan's and Zhang et al.'s methods, respectively. This shows that PSSC-core provides consistent improvements for prediction of both helix and strand content. In short, the error reduction ranges between 5–7% for helix and 7–9% for strand content prediction when compared to the state-of-the-art competing methods.

Table VI shows that the features developed based on feature selection (see feature selection results in Table VI) provided about 70% of the improvement for helix content prediction, that is, 0.4/0.6 and 0.5/0.7 for the PDBSelect25 and LinPanPlus25 datasets, respectively, while the remaining 30% of the improvement was due to the proposed new features (see PSSC-core results in Table VI). At the same time, the improvements for the strand content prediction come in about 60% due to the new features, that is, 0.2/0.8 and 0.3/0.8 for the PDBSelect25 and LinPanPlus25 datasets, respectively, and in 40% because of the feature selection. The latter, larger improvement is due to the fact that the newly proposed features were designed focusing on the strand content prediction. We also note that the Zhang et al.'s method is slightly more accurate than Lin and Pan's method for the problem that concerns low homology sequences. In contrast, the latter method was shown superior when compared with the former when sequences characterized by higher homology were considered.[12] Finally, we note that the improved accuracy of the PSSC-core method comes as a trade-off for using more features. The proposed method uses more features (46 and 88) than the methods in Refs. 11 and 12 (less than 20 per each secondary structure). At the same time, computation of all these features can be accomplished with a single pass through the protein sequence.

The statistical significance of the differences in prediction quality between the proposed method and the competing methods was measured using a paired *t*-test over the 10-fold cross validation results. The results are summarized in Table VII.

The results indicate that the differences between the PSSC-core and both competing method are statistically significant at 99.95% significance level for both datasets and secondary structures. As the standard significance level is usually set at 95% to assume that the results of a given method are significantly better than results of competing methods, we conclude with high confidence that the proposed method provides statistically significant improvements over previously reported research.

Scatter plots shown in Figure 3 for the helix content prediction and in Figure 4 for strand content prediction provide further insight into the quality of the predictions by the PSSC-core and the other two methods. The fig-

Table VII

Statistical Significance Test Between the Results of the Two Competing Method and the Proposed PSSC-Core Method for the PDBSelect25 and LinPanPlus25 Datasets; Positive t*-Value Indicates that the Proposed Method Provided More Accurate Predictions*

	Helix						Strand					
	Method in Ref. 12			Method in Ref. 11			Method in Ref. 12			Method in Ref. 11		
		Significance			Significance			Significance			Significance	
Dataset	t-*Value*	Yes/no	Level	t-*Value*	Yes/no	Level	t-*Value*	Yes/no	Level	t-*Value*	Yes/no	Level
PDBSelect25	6.21	Yes	99.95%	6.95	Yes	99.95%	10.80	Yes	99.95%	9.64	Yes	99.95%
LiPanPlus25	6.78	Yes	99.95%	6.59	Yes	99.95%	10.47	Yes	99.95%	8.22	Yes	99.95%

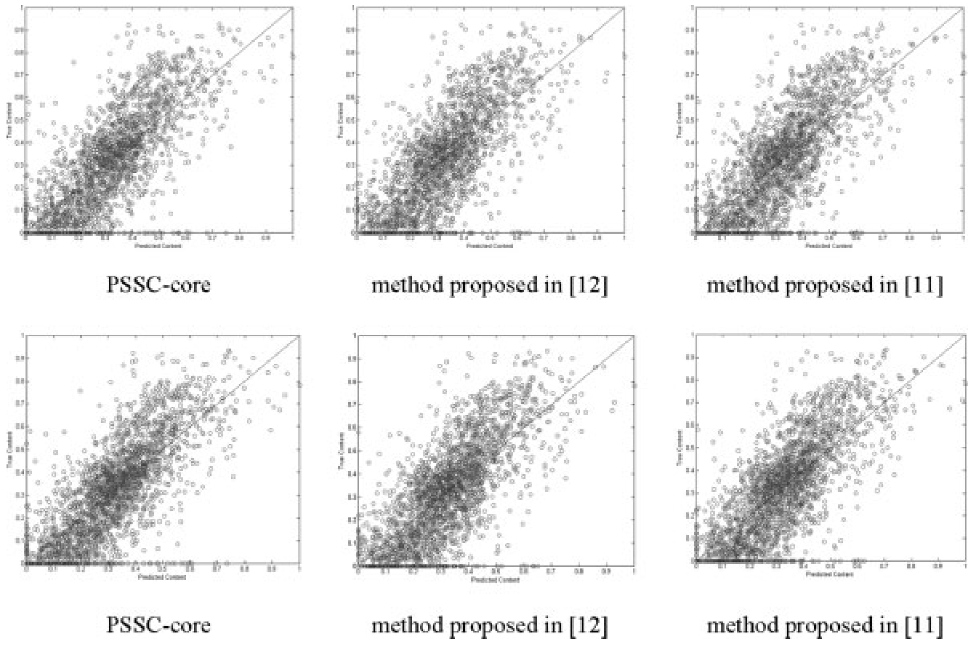
ures show results of the 10-fold cross-validation where *x*-axis corresponds to the true content values, while the *y*-axis shows the predicted content values.

The results of our method are clustered visibly closer to the diagonal, which represents perfect predictions. Similar patterns are observed for both PDBSelect25 (upper rows) and LinPanPlus25 datasets (lower rows).

CONCLUSIONS

Prediction of secondary structure content is an important and difficult computational challenge, especially when considering low homology sequences for which sequence alignment-based methods are characterized by relatively low quality predictions. The PSSC-core method, on the other hand, provides not only an accurate alternative for predicting the secondary structure content but also provides useful insight into the design of successful protein sequence representations that can be used in other methods related to prediction of different aspects of the secondary protein structure.

The PSSC-core method was shown to provide very good prediction quality for prediction of both helix and strand content for proteins in the twilight zone. Our method reduced the prediction errors by 5–7% for helix and 7–9% for strand content prediction tasks when compared with other state-of-the-art prediction methods. The improvement is mostly due to new, comprehensive and feature-based sequence representations. Two repre-



PSSC-core

method proposed in [12]

method proposed in [11]

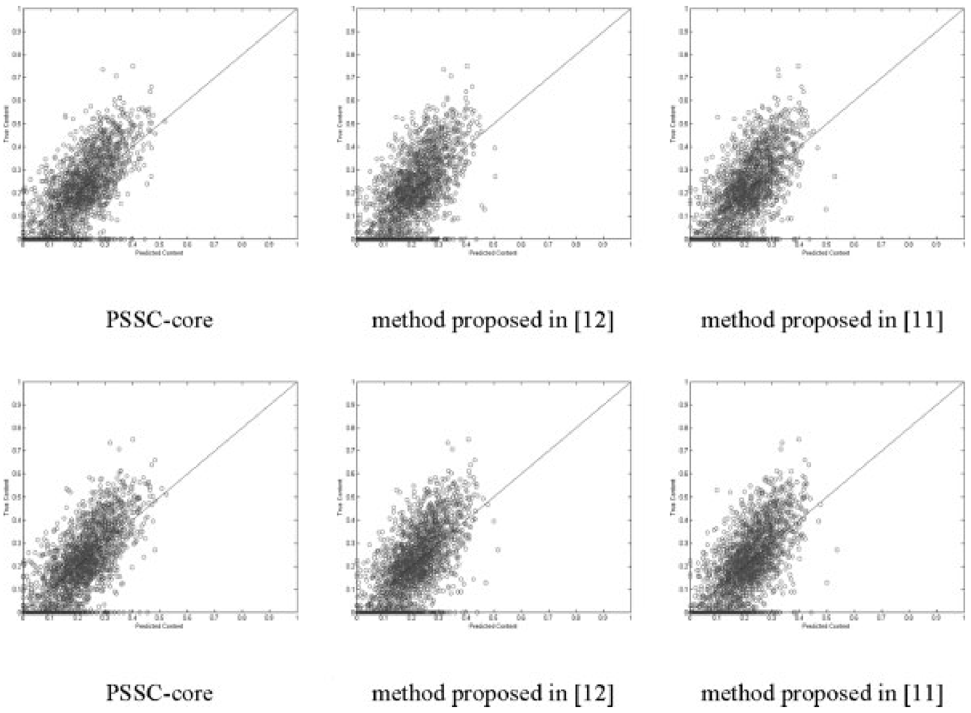
PSSC-core

method proposed in [12]

method proposed in [11]

Figure 3

Scatter plots showing the quality of the helix content prediction; the upper row shows results on the PDBSelect25 dataset, while the lower row shows results for the LinPanPlus25 dataset.



PSSC-core

method proposed in [12]

method proposed in [11]

PSSC-core

method proposed in [12]

method proposed in [11]

Figure 4

Scatter plots showing the quality of the strand content prediction; the upper row shows results on the PDBSelect25 dataset, while the lower row shows results for the LinPanPlus25 dataset.

sentations, for helix and strand content predictions, which utilize a synergic combination of features selected from the agglomerated set of features used in the past predictions of various aspect of protein structure and a set of newly designed features, were proposed. They include variety of physicochemical information concerning composition and composition moment vectors, frequency of tetra-peptides associated with helical and strand conformations, exchange groups, chemical groups of the side chains, hydrophobic group, and auto-correlation functions based on hydrophobicity, side-chain masses, hydropathy, and conformational patterns for β-sheets.

REFERENCES

1. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
2. Eyrich AV, Przybylski D, Koh IY, Grana O, Pazos F, Valencia F, Rost B. CAFASP3 in the spotlight of EVA. Proteins 2003;53(Suppl 6): 548–560.
3. Chou KC. Using pair-coupled amino acid composition to predict protein secondary structure content. J Protein Chem 1999;18:473–80.
4. Liu WM, Chou KC. Prediction of protein secondary structure content. J Protein Eng 1999;12:1041–1050.
5. Cai Y, Liu XJ, Chou KC. Prediction of protein secondary structure content by artificial neural network. J Comput Chem 2003;24:727–731.
6. Lee S, Lee BC, Kim D. Prediction of protein secondary structure content using amino acid composition and evolutionary information. Proteins 2006;62:1107–1114.
7. Krigbaum WR, Knutton SP. Prediction of the amount of secondary structure in a globular protein from its amino acid composition. Proc Natl Acad Sci USA 1973;70:2809–2813.
8. Muskal SM, Kim S-H. Predicting protein secondary structure content: a tandem neural network approach. J Mol Biol 1992;225: 713–727.
9. Eisenhaber F, Imperiale F, Argos P, Frommel C. Pediction of secondary structural contents of proteins from their amino acid composition alone, I new analytic vector decomposition methods. Proteins 1996;25:157–168.
10. Zhang CT, Lin ZS, Zhang Z, Yan M. Prediction of helix/strand content of globular proteins based on their primary sequences. Protein Eng 1998;11:971–979.
11. Zhang ZD, Sun ZR, Zhang CT. A new approach to predict the helix/strand content of globular proteins. J Theor Biol 2001;208:65–78.
12. Lin Z, Pan X. Accurate prediction of protein secondary structural content. J Protein Chem 2001;20:217–220.
13. Pilizota T, Lucic B, Trinajstic N. Use of variable selection in modeling the secondary structural content of proteins from their composition of amino acid residues. J Chem Inf Comput Sci 2004;44:113–121.
14. Ruan J, Wang K, Yang J, Kurgan LA, Cios K. Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences. Artif Intell Med 2005;35:9–35.

15. Zhang CT, Zhang Z, He Z. Prediction of the secondary structure of globular proteins based on structural classes. J Protein Chem 1996; 15:775–786.
16. Zhang CT, Zhang Z, He Z. Prediction of the secondary structure contents of globular proteins based on three structural classes. J Protein Chem 1998;17:261–272.
17. Kurgan LA, Homaeian L. Prediction of structural classes for protein sequences and domains-impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recognit 2006;39:2323–234.
18. Kedarisetti K, Kurgan LA, Dick S. Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 2006;348:981–988.
19. Sander C, Schneider R. Database of homology-derived structures and the structural meaning of sequence alignment. Proteins 1991;9: 56–68.
20. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
21. Pollastri G, McLysaght A. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 2005;21:1719–1720.
22. Lin K, Simossis VA, Taylor WR, Heringa J. A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics 2005;21:152–159.
23. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522.
24. Chandonia JM, Karplus M. New methods for accurate prediction of protein secondary structure. Proteins 1999;35:293–306.
25. Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics 2002;18:77–82.
26. Homaeian L. Towards improving accuracy of protein content prediction for low homology sequences. MSc Thesis, Department of Electrical and Computer Engineering, University of Alberta, 2006.
27. Hocking RR. Methods and applications of linear models: regression and analysis of variance (Wiley Series in Probability and Statistics). New York: Wiley; 1996.
28. Hobohm U, Sander C. A sequence property approach to searching protein databases. J Mol Biol 1995;251:390–399.
29. Wang J, Ma Q, Shasha D, Wu CH. Application of neural networks to biological data mining: a case study in protein sequence classification. In: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining, 2000. pp 305–309.
30. Luo R, Feng Z, Liu J. Prediction of protein structural class by amino acid and polypeptide composition. Eur J Biochem 2002;269: 4219–4225.
31. Syed U, Yona G. Using a mixture of probabilistic decision trees for direct prediction of protein function. In: Proceedings of RECOMB, 2003 conference, 2003. pp 224–234.
32. Yang X, Wang B. Weave amino acid sequences for protein secondary structure prediction. In: Proceedings of the eight ACM SIGMOD workshop on research issues in data mining and knowledge discovery, 2003. pp 80–87.
33. Ganapathiraju MK, Klein-Seetharaman J, Balakrishnan N, Reddy R. Characterization of protein secondary structure. IEEE Signal Process Mag 2004;15:78–87.
34. Cao Y, Liu S, Zhang L, Qin J, Wang J, Tang K. Prediction of protein structural class with rough sets. BMC Bioinformatics 2006;7:20.
35. Sun X-D, Huang RB. Prediction of protein structural classes using support vector machines. Amino Acids 2006;30:469–475.
36. Kurgan LA, Homaeian L. Prediction of secondary protein structure content from primary sequence alone – a feature selection based approach. In Proceedings of the international conference on machine learning and data mining in pattern recognition, Leipzig, Germany, 2005. pp 334–345.
37. Sweet R, Eisenberg D. Correlation of sequence hydrophobicities measures similarity in three dimensional protein structure. J Mol Biol 1983;171:479–488.
38. Li X, Pan X. New method for accurate prediction of solvent accessibility from protein sequence. Proteins 2001;42:1–5.
39. Cai Y, Liu XJ, Xu XB, Chou KC. Support vector machines for prediction of protein domain structural class. J Theor Biol 2003;221: 115–120.
40. Chou KC, Cai YD. Prediction protein structural class by functional domain composition. Biochem Biophys Res Commun 2004;321: 1007–1009.
41. Nelson D, Cox M. Lehninger principles of biochemistry amino. New York: Worth Publishers; 2000.
42. Fauchere JL, Pliska V. Hydrophobic parameters p of amino-acid side chains from the partitioning of *N*-acetyl-amino-acid amides. Eur J Med Chem 1983;18:369–375.
43. Eisenberg D, Weiss RM, Trewilliger TC. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci USA 1984;81:140–144.
44. Kyte J, Doolitle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.

DOI 10.1002/prot