Amino Acids (2008) 35:551–564
DOI 10.1007/s00726-008-0080-3

ORIGINAL ARTICLE

Secondary structure-based assignment of the protein structural classes

Lukasz A. Kurgan · Tuo Zhang · Hua Zhang ·
Shiyi Shen · Jishou Ruan

Received: 31 January 2008 / Accepted: 27 February 2008 / Published online: 22 April 2008

Abstract Structural class categorizes proteins based on the amount and arrangement of the constituent secondary structures. The knowledge of structural classes is applied in numerous important predictive tasks that address structural and functional features of proteins. We propose novel structural class assignment methods that use one-dimensional (1D) secondary structure as the input. The methods are designed based on a large set of low-identity sequences for which secondary structure is predicted from their sequence (PSSA^{sc} model) or assigned based on their tertiary structure (SSA^{sc}). The secondary structure is encoded using a comprehensive set of features describing count, content, and size of secondary structure segments, which are fed into a small decision tree that uses ten features to perform the assignment. The proposed models were compared against seven secondary structure-based and ten sequence-based structural class predictors. Using the 1D secondary structure, SSA^{sc} and PSSA^{sc} can assign proteins to the four main structural classes, while the existing secondary structure-based assignment methods can predict only three classes. Empirical evaluation shows that the proposed models are quite promising. Using the structure-based assignment performed in SCOP (structural classification of proteins) as the golden standard, the accuracy of SSA^{sc} and PSSA^{sc} equals 76 and 75%, respectively. We show that the use of the secondary structure predicted from the sequence as an input does not have a detrimental effect on the quality of structural class assignment when compared with using secondary structure derived from tertiary structure. Therefore, PSSA^{sc} can be used to perform the automated assignment of structural classes based on the sequences.

Keywords SCOP · Structural class · Structural class prediction · Structural classification of proteins · Secondary protein structure

Introduction

Structural class constitutes one of the coarsest structural classifications of proteins in which protein structures are categorized based on the amounts and arrangement of the constituent secondary structures. Despite the low granularity of this categorization, a knowledge of structural classes provides useful input for a variety of important applications, including the prediction of a variety of functional and structural properties and the determination of distant homologues. The structural class is currently assigned manually based on the tertiary structure, which is the reason why this information is known for only a relatively small number of proteins. The recent release (1.73) of the SCOP (structural classification of proteins) database (Murzin et al. 1995), which includes annotation of structural classes, includes 97,178 protein domains. At the same time, release 27 of the NCBI's RefSeq database includes 4,426,609 non-redundant protein sequences. The main reason for such a wide gap is the unavailability of protein

L. A. Kurgan (✉) · T. Zhang · H. Zhang
Department of Electrical and Computer Engineering,
University of Alberta, 2nd floor, ECERF (9107 116 Street),
T6G 2V4 Edmonton, AB, Canada
e-mail: lkurgan@ece.ualberta.ca

T. Zhang · H. Zhang · S. Shen · J. Ruan
College of Mathematical Science and LPMC, Nankai University,
Tianjin, People's Republic of China

S. Shen · J. Ruan
Chern Institute of Mathematics, Tianjin,
People's Republic of China

structure, which is used to assign the protein to the corresponding structural class, for the significant majority of known protein sequences. Therefore, an accurate, automated method for classifying sequences into the corresponding structural classes would provide the needed help in the laborious task of populating the SCOP database. To this end, we address the following three aims:

1. We investigate whether the structural class can be successfully assigned/predicted based on one-dimensional (1D) secondary structure, i.e., secondary structure assigned to each residue without the knowledge of its spatial arrangement.
2. We also analyze whether the early definitions of structural classes, which were developed before the SCOP database was established, can be successfully used to automate the assignment of the classes based on the 1D secondary structure.
3. The above goals are addressed using the assigned (based on the tertiary structure) and the predicted (from protein sequence) 1D secondary structure. We propose two new assignment models: SSA^{sc} (secondary structure-based assignment of structural classes) and PSSA^{sc} (predicted secondary structure-based assignment of structural classes). We compare these two models against existing assignment methods (goal 2) to verify whether the structural class can be successfully assigned using 1D secondary structure predicted from the protein sequence. Such an assignment model would constitute a feasible approach to automate the assignment of structural classes without the knowledge of the structure.

We also compare the above results against representative methods that perform prediction of structural classes based on the protein sequence.

This paper is organized as follows. We first discuss the definitions and applications of structural classes and introduce methods currently being used for sequence-based prediction of structural classes. Next, we introduce the datasets, data, and methods used to address the above-mentioned goals. This is followed by the discussion of the proposed models for structural class assignment. Finally, we present, analyze, and compare the results from out empirical evaluation of the proposed models with other assignment and prediction methods, and we draw conclusions with respect to the defined goals.

Definitions of structural classes

The concept of structural classes was first proposed by Levitt and Chothia in the mid-1970s (Levitt and Chothia 1976). In their pioneering work, these researchers evaluated and classified 31 structures of globular proteins into four structural classes: (1) all-α class, which includes proteins with only small amount of strands; (2) all-β class with proteins, with only a small amount of helices; (3) the α/β class with proteins, which include both helices and strands and where the strands are mostly parallel; (4) $\alpha + \beta$ class, which includes proteins with both helices and strands and where the strands are mostly antiparallel. These definitions were later modified and made more specific with respect to the required amounts of helices and strands (see Table 1). All of these classifications also consider irregular proteins (also called ξ proteins) (Chou and Zhang 1993), i.e., proteins which can not be assigned to one of the four structural classes, which are rare and therefore usually omitted from predictions.

The common feature of the assignment methods listed in Table 1 is that the structural classes are defined based on the helical and strand content as well as the orientation of the β-sheets. The main differences lie in the threshold values used to define minimal/maximal amounts of strands and helices for a given structural class.

In 1986 Nakashima and colleagues defined five structural classes using 135 tertiary protein structures (Nakashima et al. 1986). This was followed by Klein and DeLisi who in their definitions enlarged the space covered by the irregular proteins (Klein and DeLisi 1986). The next definition was proposed by P.Y. Chou in 1989, who further enlarged the set of irregular proteins by increasing the threshold values for the four main structural classes (Chou 1989). The definition proposed by Kneller and colleagues lowered the thresholds for the $\alpha + \beta$ and α/β classes (Kneller et al. 1990). This was followed in 1995 by K.C. Chou who used 129 proteins to propose another classification into five classes (Chou 1995). The definitions by K.C. Chou are based on secondary structure content defined using the Dictionary of Secondary Structure of Proteins (DSSP, Kabsch and Sander 1983). The next definition, which assumes four structural classes by combining the $\alpha + \beta$ and α/β classes into the so-called mixed class, was proposed by Eisenhaber and colleagues in 1996 (Eisenhaber et al. 1996). The reason to consider a mixed class was that the authors proposed their definition based solely on 1D secondary structure assigned with DSSP, in which case information about strand directionality was unavailable. Their definition is equivalent to the definition proposed by Nakashima and colleagues. Finally, in 1998, Liu and Chou refined their assignment rules originally proposed in Chou (1995) to increase the size of the regions associated with the four structural classes and to improve the definitions of the $\alpha + \beta$ and α/β classes (Liu and Chou 1998).

The threshold-based classifications were deemed obsolete in the late 1990s and replaced by the manually preformed SCOP classification. The SCOP database includes a description of the structural and evolutionary relationships of proteins from the Protein Data Bank (PDB)

Table 1 Structural class definitions

Reference's	Structural class	Helix (α) amount	Strand (β) amount	Additional constrains and comments
Nakashima et al. (1986)	α proteins	>15%	<10%	
	β proteins	<15%	>10%	
	$\alpha + \beta$ proteins	>15%	>10%	Contains dominantly antiparallel β-sheets
	α/β proteins	>15%	>10%	Contains dominantly parallel β-sheets
	ξ proteins	<15%	<10%	
Klein and DeLisi (1986)	α proteins	>40%	<5%	
	β proteins	<10%	>30%	
	$\alpha + \beta$ proteins	≥15%	≥15%	Contains dominantly antiparallel β-sheets
	α/β proteins	≥15%	≥15%	Contains dominantly parallel β-sheets
	ξ proteins			$\alpha + \beta < 20\%$
Chou (1989)	α proteins	>45%	<5%	
	β proteins	<5%	>45%	
	$\alpha + \beta$ proteins	>30%	>20%	Contains dominantly antiparallel β-sheets
	α/β proteins	>30%	>20%	Contains dominantly parallel β-sheets
	ξ proteins			Otherwise
Kneller et al. (1990)	α proteins	≥30%	$\leq 0.15(\alpha + \beta)$	
	β proteins	<10%		
	$\alpha + \beta$ proteins	>15%	>5%	Contains dominantly antiparallel β-sheets
	α/β proteins	>15%	>5%	Contains dominantly parallel β-sheets
	ξ proteins			Otherwise
Chou (1995)	α proteins	≥40%	≤5%	
	β proteins	≤5%	≥40%	
	$\alpha + \beta$ proteins	≥15%	≥15%	More than 60% antiparallel β-sheets
	α/β proteins	≥15%	≥15%	More than 60% parallel β-sheets
	ξ proteins	≤10%	≤10%	
Eisenhaber et al. (1996)	α proteins	>15%	<10%	
	β proteins	<15%	>10%	
	mixed proteins	>15%	>10%	No division into $\alpha + \beta$ and α/β classes
	ξ proteins			Otherwise
Liu and Chou (1998)	α proteins	≥20%	≤5%	And $\alpha - 4\beta \geq 0.2$
	β proteins	≤5%	≥20%	And $\beta - 4\alpha \geq 0.2$
	$\alpha + \beta$ proteins	≥12%	≥12%	And $\alpha + \beta \geq 30\%$ and 70% or more antiparallel bridges
	α/β proteins	≥12%	≥12%	And $\alpha + \beta \geq 30\%$ and 60% or fewer antiparallel bridges
	ξ proteins	≤10%	≤10%	
SCOP Murzin et al. (1995)	α proteins	N/A	N/A	Manual classification
	β proteins			
	$\alpha + \beta$ proteins			
	α/β proteins			
	+7 other classes			

(Berman et al. 2000). The SCOP database classifies proteins on multiple levels, including structural classes, but also as belonging to different families and super-families and containing different domains. The SCOP's classification does not incorporate hardcoded rules for structural classes. Instead, the decisions are made based on structural elements that are located in individual domains that constitute the protein. Researchers claim that the SCOP classification is more "natural" and provides more reliable information to study protein structural classes when compared to classifications based on the percentage amounts of the secondary structures (Murzin et al. 1995; Chou and Maggiora 1998; Wang and Yuan 2000). The SCOP classification currently includes 11 classes (Andreeva et al. 2004): (1) all-α proteins; (2) all-β proteins; (3) α/β proteins; (4) $\alpha + \beta$ proteins; (5) multi-domain proteins; (6) membrane and cell surface proteins; (7) small proteins; (8) coiled coils proteins; (9) low-resolution proteins; (10) peptides; (11) designed proteins. Only the first four categories are usually considered for computational prediction purposes as they cover a significant majority of the proteins.

Applications of structural classes

Information on the structural classes of proteins is useful for studying the broader problem of protein structure prediction and carrying out a number of predictive tasks that address certain structural and functional features. More specifically, a knowledge of structural classes has been applied to improve the accuracy of secondary structure prediction (Gromiha and Selvaraj 1998), to reduce the search space of possible conformations of the tertiary structure (Chou 1992; Chou and Zhang 1995; Bahar et al. 1997), and to implement a heuristic approach to determine tertiary structure (Carlacci et al. 1991). Information on structural classes has also been used to provide useful input for numerous predictive applications that include the discrimination of outer membrane proteins (Gromiha 2005a; Gromiha and Suwa 2005) and the prediction of protein-folding rates (Gromiha 2005b) and -unfolding rates (Gromiha et al. 2006), DNA-binding sites (Kuznetsov et al. 2006), protein folds (He et al. 2002), and secondary structure content (Zhang et al. 1998, 2001). In particular, the concepts and algorithms developed in protein structural class prediction have greatly stimulated the development of predicting many other protein attributes (Chou 2005b), such as subcellular localization (Cedano et al. 1997; Chou and Elrod 1999; Chou and Shen 2007b, 2008; Xiao et al. 2005, 2006b), membrane protein type (Chou and Shen 2007a, c), enzyme functional class (Shen and Chou 2007a), GPCR type (Chou 2005a; Wen et al. 2007), and signal peptides (Chou and Shen 2007b, c). The wide range of the applications and quality of venues in which these applications were published provide strong evidence supporting the usefulness of the structural classes in many aspects of protein research.

Prediction of structural classes from the protein sequence

Since the manual assignment of structural classes performed in SCOP cannot be directly traced using the input protein sequence or even its corresponding secondary structure, a variety of methods that predict the structural class based on the protein sequence have been developed to facilitate an automated, high-throughput assignment.

Prediction of the structural classes is performed in two steps: (1) the AA sequences are transformed into fixed-length feature vectors; (2) the feature vectors are inputted into a classification model to generate the corresponding structural class. The majority of existing structural class prediction methods use relatively simple feature vectors that incorporate composition vector, auto-correlation functions based on non-bonded residue energy, polypeptide composition, pseudo AA composition, and complexity measure factors (Chou and Zhang 1994, 1995; Chou 1995; Chou and Maggiora 1998; Chou et al. 1998; Zhou 1998; Jin et al. 2003; Cai et al. 2006; Kedarisetti et al. 2006a; Xiao et al. 2006a; Chen et al. 2008). Some recent methods use hybrid feature vectors that combine physicochemical properties and sequence composition, while others optimize one selected feature (Du et al. 2006; Kedarisetti et al. 2006b; Jahandideh et al. 2007; Kurgan and Homaeian 2006; Kurgan and Chen 2007). The predictions are performed using a variety of classification algorithms that include fuzzy clustering (Shen et al. 2005), fuzzy nearest neighbor (Zhang et al. 2008), neural network (Cai and Zhou 2000; Cai et al. 2002b), Bayesian classification (Wang and Yuan 2000), rough sets (Cao et al. 2006), component-coupled (Chou and Maggiora 1998; Chou et al. 1998; Zhou 1998), information discrepancy (Jin et al. 2003; Kedarisetti et al. 2006a), logistic regression (Kedarisetti et al. 2006b; Kurgan and Homaeian 2006; Jahandideh et al. 2007; Kurgan and Chen 2007), decision trees (Cai et al. 2006; Dong et al. 2006), and support vector machine (Cai et al. 2001, 2002a, 2003; Dong et al. 2006; Kedarisetti et al. 2006b; Sun and Huang 2006; Zhang and Ding 2007). Recent studies have used multi-classifier models, such as ensembles (Kedarisetti et al. 2006b), bagging (Dong et al. 2006), and boosting (Feng et al. 2005; Cai et al. 2006; Niu et al. 2006). An in-depth review of computational methods used for predicting structural classes can be found in Chou (2005b). One feasible approach to obtaining accurate structural class prediction is to use a large library of reference functional sequence motifs to build a feature vector that is subsequently used as the input to the classification algorithm. Such a method was proposed by Chou and Cai (2004): 7785 features were used, and a 98% accuracy on a set of proteins characterized by low-sequence identity was obtained for seven structural classes.

In this article, however, we investigate whether the 1D secondary structure (either the secondary structure assigned using DSSP or the secondary structure predicted from the protein sequence) can be used to predict/assign the structural class.

Materials and methods

Dataset

The goals defined in this paper were investigated using a large benchmark dataset of representative twilight zone protein sequences. The dataset, called 25PDB, was selected using the 25% PDBSELECT list compiled in July 2005 (Hobohm and Sander 1994), which includes proteins from PDB that were scanned with high-resolution and which are characterized by low (on average about 25%) sequence identity. The dataset was originally published in Kurgan and Homaeian (2006) and has been used to benchmark two recent structural class prediction methods (Kedarisetti et al. 2006b; Kurgan and Chen 2007). It contains 1673 proteins and domains, which include 443 all-α, 443 all-β, 346 α/β, and 441 $\alpha + \beta$ sequences.

Predicted versus assigned secondary structure

The secondary structure, which constitutes the input for the assignment/prediction methods studied in this paper, can be obtained from two sources:

- It can be extracted from the tertiary structure (atomic coordinates) stored in PDB. Although there are numerous applications that can perform such secondary structure assignment (Martin et al. 2005), the most popular assignment method is the DSSP (Kabsch and Sander 1983), which is the method used to annotate proteins stored in PDB. The DSSP defines eight types of secondary structures that are combined into three main secondary structure states: helix, strand, and coil.
- It can be predicted from the protein sequence. The advantage of these prediction methods is that they do not require knowledge of the underlying tertiary structure. At the same time, they only obtain about 78–80% accuracy with respect to the actual secondary structure assignment performed with DSSP (Lin et al. 2005; Birzele and Kramer 2006). We use the PSI-PRED secondary structure prediction method (Jones 1999; Bryson et al. 2005) for two reasons: (1) it has recently been shown to provide superior accuracy in comparison with other state-of-the-art secondary structure prediction methods (Lin et al. 2005; Birzele and Kramer 2006); (2) this method is frequently used to support a variety of other predictions tasks, such as fold prediction (Chen and Kurgan 2007), folding rate prediction (Ivankov and Finkelstein 2004), and the prediction of β-turns (Fuchs and Alix 2005), α-turns (Wang et al. 2006), solvent accessibility (Garg et al. 2005), contact orders (Song and Burrage 2006), and disulfide connectivity (Song et al. 2007), among others.

Feature-based representation of secondary structure

As with sequence-based prediction methods, the proposed approach requires two steps. The main difference between the two methods is that the input in the latter is the secondary structure (either assigned with DSSP or predicted with PSI-PRED) rather than the sequence. Although the secondary structure content that is used by the existing structural class assignment methods (see Table 1) reflects information about the secondary structure of the entire sequence, it does not provide information on individual secondary structure segments. The size (length) of the secondary structure segments is one of the important factors when it comes to the assignment of the structural classes in SCOP. In designing our features, we assert that although the secondary structure prediction accuracy is only about 80%, the predicted secondary structure preserves enough information about the secondary structure segments to characterize the structural class. To this end, we developed the following set of features to represent the secondary structure for the purpose of the structural class assignment/prediction:

- composition moment vector (CMV)

$$\mathrm{CMV}_I^k = \frac{\sum_{j=1}^{n_I} n_{Ij}^k}{\prod_{d=0}^{k} (N-d)}$$

  where I = [helix (h), strand (e), coil (c)], n_{Ij} is the jth position (in the secondary structure sequence) of the Ith secondary structure type, n_I is the total number of residues having Ith secondary structure, N is the length of the protein sequence, and k is the order of the CMV. We apply CMVs for $k = 0,1,\ldots,5$. CMV_1^0 reduces to the secondary structure content, which is used as the input for the existing secondary structure based assignment methods.
- count of secondary structure segments that include at least k residues for coil segments

$$\mathrm{Count}L_c^k = \frac{\sum_{j=k}^{20} \mathrm{count}_c^j}{\sum_{I=\{h,e,c\}} \mathrm{total}_I} \quad \text{for} \quad k = 2, 3, \ldots 20$$

  for helix segments

$$\mathrm{Count}L_h^k = \frac{\sum_{j=k}^{20} \mathrm{count}_h^j}{\sum_{I=\{h,e\}} \mathrm{total}_I} \quad \text{for} \quad k = 3, 4, \ldots 20$$

  for strand segments

$$\mathrm{Count}L_e^k = \frac{\sum_{j=k}^{20} \mathrm{count}_e^j}{\sum_{I=\{h,e\}} \mathrm{total}_I} \quad \text{for} \quad k = 2, 3, \ldots 20$$

  where i = [helix (h), strand (e), coil (c)], $\mathrm{count}_{c,\,h,\,e}^j$ denotes the number of coil, strand, and helix segments of length j in the predicted/assigned secondary structure, and total_I denotes the total number of all segments belonging to Ith secondary structure. The smallest α-helix segment is assumed to include at least three residues. The upper bound on the segment size is set to 20 since larger segments occur rarely. The count of coil segments is normalized by the total number of all segments, while the counts of strand and helix segments are normalized by the total number of strand and helix segments. These differences in the normalizations accommodate all-α and all-β classes that may not include any strand and helix segments, respectively.

- average and maximal size of secondary structure segments
  - length of the longest segment, in terms of

    MaxSeg_I where $I = \{\text{helix}(h), \text{strand}(e), \text{coil}(c)\}$
  - normalized length of the longest segment

    $N\text{MaxSeg}_\text{i} = \text{MaxSeg}_I/N$ where $I = \{\text{helix}(h), \text{strand}(e), \text{coil}(c)\}$
  - average length of the segment

    AvgSeg_I where $I = \{\text{helix}(h), \text{strand}(e), \text{coil}(c)\}$
  - normalized average length of the segment

    $N\text{AvgSeg}_\text{i} = \text{AvgSeg}_I/N$ where $I = \{\text{helix}(h), \text{strand}(e), \text{coil}(c)\}$

There are a total of 86 generated features, including 18 CMV features, 56 normalized counts, and 12 features that describe the longest and average segment lengths.

Classification model

Once the input secondary structure is converted into the feature vector, this information is inputted to compute the model for prediction/assignment of the structural classes. Our aim is to provide an easy to comprehend model that can be contrasted against the existing assignment methods. Although a wide range of different computational techniques can be used to derive the model, most of the resulting models are very complex and, therefore, incomprehensible to humans. Examples of such models are neural networks, Bayesian models, information discrepancy-based models, and support vector machines. The remaining techniques, which include decision trees, regression, and rule-based models, have the advantage of being expressed in a human-readable format. Among these techniques, we chose to use decision trees as they provide a model that is easy to comprehend and analyze, and they have been successfully used in sequence-based prediction of structural classes (Cai et al. 2006; Dong et al. 2006). We employed the divide-and-conquer decision tree algorithm C4.5 (Quinlan 1993) revision 8 implemented in the software package WEKA ver. 3.4.6 (Witten and Frank 2005).

Results and discussion

Experimental setup

The defined goals are investigated with the use of the 25PDB dataset and two experimental test types:

- The resubstitution test in which the structural class assignment model is generated using the entire 25PDB set and then the model is tested on the same dataset. This test is used to compare our assignment models with existing secondary structure-based methods for assignment of structural classes.
- The tenfold cross validation (10CV) test randomly divides the 25PDB set into ten subsets, and ten assignment models are generated. Each time one of the ten subsets is used to test the model and the remaining nine are used to generate it. We use this test to compare our models against a selected set of representative sequence-based structural class prediction methods.

We chose to perform 10CV due to its favorable computational cost when compared with the jackknife cross-validation test. At the same time, the jackknife test is considered to be more objective since it always yields the same results for a given benchmark dataset while the sub-sampling (such as fivefold or tenfold) cross-validation may provide arbitrary test results due to large number of potential sub-samples, as shown in Chou and Shen (2007b, Chou and Shen 2008).

The resubstitution-based model is computed using the same parameters of the C4.5 algorithm as the 10CV models. A separate set of tests is performed when using DSSP-assigned and PSI-PRED-predicted secondary structure as the input data. The results reported here include the overall accuracy (the number of correct predictions divided by the total number of test sequences) and accuracy for each structural class (number of correct predictions for a given class divided by the number of sequences in this class).

Structural class assignment models

The 25PDB dataset encoded using 86 features was inputted into the C4.5 decision tree algorithm to generate structural class assignment models, one for input data based on DSSP-assigned secondary structure (SSA^{sc}) and another using secondary structure predicted with PSI-PRED (PSSA^{sc}). The SSA^{sc} model is shown in Fig. 1.

The SSA^{sc} model includes 12 rules that allow a given secondary structure sequence to be assigned to one of the four structural classes: one rule for the all-α class; three rules for the all-β class; two rules for the α/β class, and six

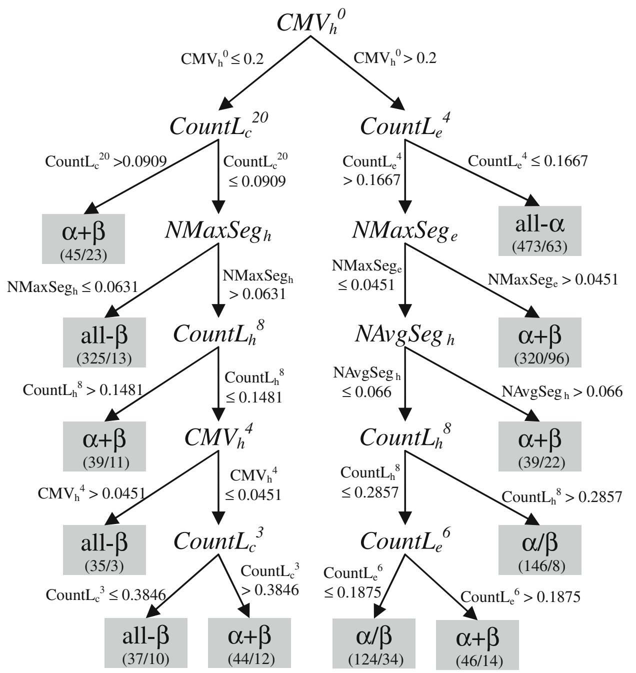


Fig. 1 The secondary structure-based assignment of structural classes (SSA^{sc}) model generated from secondary structure assigned with the Dictionary of Secondary Structure of Proteins (DSSP) using the 25PDB dataset

rules for the $\alpha + \beta$ class. A rule is assembled by following a path through a decision tree starting at the top node (root) and following through the branches, which corresponds to conditions in a rule, to reach a leaf node (denoted using shaded box in Fig. 1) that defines the outcome. For example, the rule for all-α class reads as follows:

$$\text{IF}\,\text{CMV}_h^0 > 0.2 \quad \text{and} \quad \text{Count}L_e^4 \leq 0.1667\,\text{THEN all - }\alpha$$

The leaf nodes also show the success/failure rate of a given rule for the 25PDB dataset, i.e., the above rule for the all-α class results in 473 correct assignments into this class and 63 incorrect assignments (proteins from other classes that are classified as all-α class).

The PSSA^{sc} model includes 12 rules: four, one, two, and five rules for the all-α, all-β, α/β, and $\alpha + \beta$ classes, respectively (Fig. 2).

As in the the existing assignment methods, the helix and strand content were used in our models. At the same time, a number of other features were utilized, mostly to differentiate between the α/β and $\alpha + \beta$ classes. The two models use only a small subset of ten features from the original set of 86 features (Table 2). One of the advantages of the C4.5 algorithm is that it selects the most discriminative features to build the classification model and disregards the remaining features. This means that the proposed structural class assignment requires only ten features, although we note that they come from each of the proposed feature categories. Both models use the helix content (CMV$_h^0$), CountL_h^8, NMaxSeg$_h$, and NMaxSeg$_e$ features, and some of the remaining features used in only one of the models are also similar, including CountL_e^6 versus CountL_e^7 or CMV$_h^4$ versus CMV$_h^3$. This is expected as both models have the same goal. At the same time, some other features, such as strand content (CMV$_e^0$), counts of coil segments, and average length of helical segments, are used only in one of the models. In general, the assignment is performed based on the knowledge of all three secondary structures (including coils), and knowledge on both the structural content and the segment size is used.

Comparison with existing structural class assignment methods

The proposed structural class assignment methods are compared with seven existing assignment methods. Although the proposed models are capable of predicting $\alpha + \beta$ and α/β classes directly from the 1D secondary structure, the remaining methods combine these classes into the mixed class. The reason for this is because the models require knowledge of the directionality of the β-sheets to differentiate between $\alpha + \beta$ and α/β classes, and this information is not available in the 1D secondary structure. Therefore, for the purpose of this test we combine our $\alpha + \beta$ and α/β assignments into the mixed class. We also assume that any protein for which none of the conditions from Table 1 for a given assignment method is satisfied is automatically assigned into the irregular class. The comparison of the quality of the assignment of structural classes by the two proposed and the seven existing methods (the existing methods are tested with both DSSP-assigned and PSI-PRED-predicted secondary structure) on the 25PDB dataset are shown in Table 3.

Several interesting conclusions can be drawn based on these results:

- The use of the predicted secondary structure does not have a detrimental effect on the quality of the structural class assignment. In some cases, for example, where methods proposed in Klein and DeLisi (1986), Chou (1989), Kneller et al. (1990), Chou (1995), Liu and Chou 1998, the results are even better when the predicted secondary structure is used. This is an important observation since it means that the structural class can be effectively assigned on the sole basis of the sequence, i.e., the PSI-PRED uses only the sequence to predict the secondary structure.
- The most accurate existing assignment method was proposed in Nakashima et al. (1986) and later reused in Eisenhaber et al. (1996). The helix and strand content thresholds set in this method equal 15 and 10%,

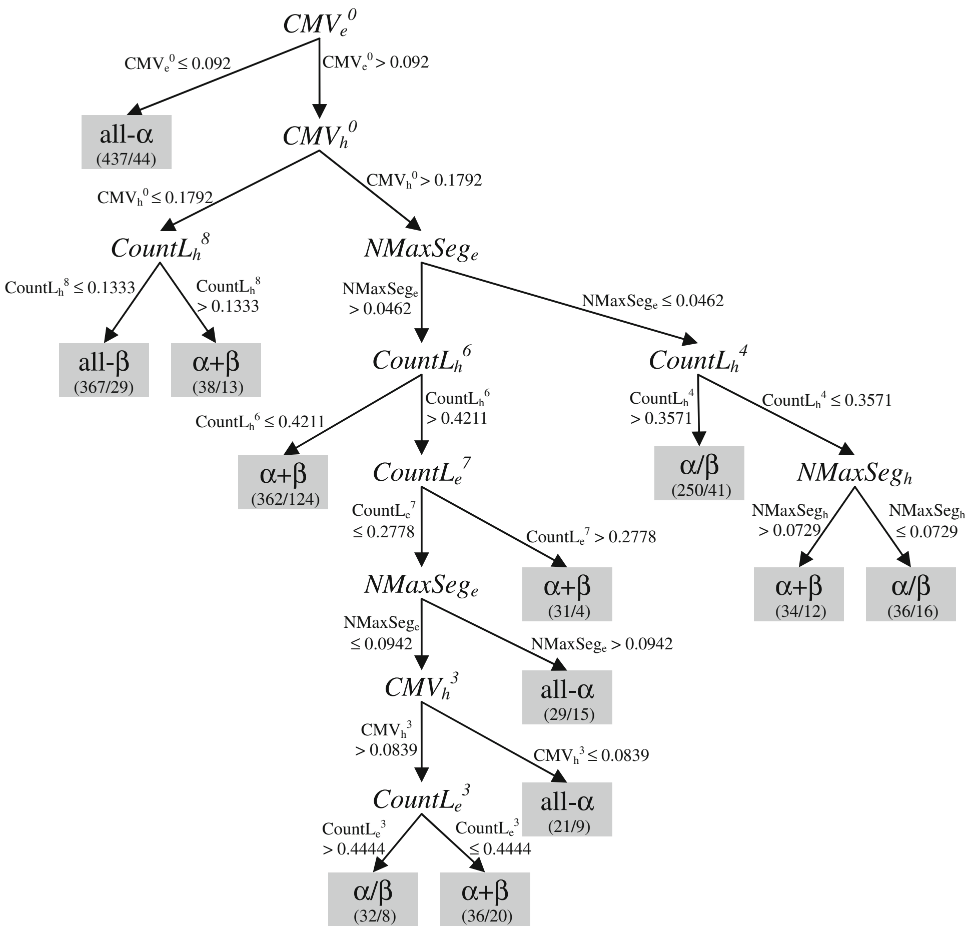


Fig. 2 The predicted secondary structure-based assignment of structural classes (PSSA^{sc}) model generated from secondary structure predicted with PSI-PRED using the 25PDB dataset

Table 2 Comparison of features used in the SSA^{sc} and PSSA^{sc} models

Feature set category	SSA^{sc} model (based on DSSP)	PSSA^{sc} model (based on PSI-PRED)
Composition moment vector	CMV$_h^0$ (helix content), CMV$_h^4$	CMV$_h^0$ (helix content), CMV$_e^0$ (strand content), CMV$_h^3$
Count of secondary structure segments	CountL_h^8 CountL_e^4, CountL_e^6 CountL_c^3, CountL_c^{20}	CountL_h^4, CountL_h^6, CountL_h^8 CountL_e^3, CountL_e^7
Average and maximal size of secondary structure segments	NMaxSeg$_h$, NMaxSeg$_e$ NAvgSeg$_h$	NMaxSeg$_h$, NMaxSeg$_e$
Total number of features	10	10

SSA^{sc}, Secondary structure-based assignment of structural classes; PSSA^{sc}, predicted secondary structure-based assignment of structural classes; DSSP, Dictionary of Secondary Structure of Proteins

respectively, which approximately coincides with the extrema on the plot of distributions of these two structures in the 25PDB dataset (see Fig. 3). The same finding was also shown in Eisenhaber et al. (1996) for a different and smaller set of proteins. In contrast, other assignment methods use relatively high threshold values for helix content, i.e., between 30 and 45% for the all-α class, and for strand content, i.e., between 30 and 45% for the all-β class. As a result, their performance for these two classes is weaker. The distributions also show that the number of residues in the helical conformation is, on average, higher than the corresponding strand content, i.e., the corresponding extremes are at higher values of content. Therefore, the thresholds for the two structures should differ accordingly, which is in contrast with symmetrical thresholds

Table 3 Comparison of the accuracy of the structural class assignment for the three classes between the proposed SSA^{sc} and PSSA^{sc} methods and seven existing assignment methods

Assignment method	Source of secondary structure	Accuracy (%)			
		All-α	All-β	Mixed	Overall
Nakashima et al. (1986); Eisenhaber et al. (1996)	DSSP	91.2	77.9	86.8	85.6
Klein and DeLisi (1986)	DSSP	75.2	56.4	75.6	70.4
Chou (1989)	DSSP	71.6	18.3	25.7	35.9
Kneller et al. (1990)	DSSP	79.2	63.0	88.7	79.4
Chou (1995)	DSSP	75.2	26.0	75.6	62.3
Liu and Chou (1998)	DSSP	80.4	36.3	85.4	71.1
SSA^{sc} (this paper)	DSSP	92.6	83.7	89.5	88.8 (3 classes)
				(65.9 α/β, 80.5 α + β)	81.5 (4 classes)
Nakashima et al. (1986); Eisenhaber et al. (1996)	PSI-PRED	88.7	74.0	90.5	85.7
Klein and DeLisi (1986)	PSI-PRED	77.2	58.5	74.7	71.1
Chou (1989)	PSI-PRED	76.1	26.4	28.6	40.6
Kneller et al. (1990)	PSI-PRED	78.8	61.9	91.5	80.3
Chou (1995)	PSI-PRED	77.4	34.3	74.7	64.7
Liu and Chou (1998)	PSI-PRED	79.0	39.5	89.3	73.4
PSSA^{sc} (this paper)	PSI-PRED	94.6	76.3	89.5	87.3 (3 classes)
				(73.1 α/β, 74.4 α + β)	80.0 (4 classes)

used in method by Chou (1989), Chou (1995), and Liu and Chou (1998).

- The highest accuracies are achieved for the all-α class, while the lowest is obtained for the all-β class. This result could be explained as being due to a larger overlap of content values between the all-β class and the mixed class in comparison with the overlap between the all-α class and the mixed class; see Fig. 4. At the same time, the overlap between the all-α and all-β classes is very small, thus making it easier to distinguish between the all-α class and the other two classes based on the helix and strand content values.
- Figure 4 shows a substantial overlap between the α/β and α + β classes, which means that traditional assignment methods that are based solely on the helix and strand content would have difficulty in distinguishing between these two classes. We note that, in the case of SCOP, the α/β and α + β classes contain both α-helices and β-strands, which are mainly interspersed and segregated, respectively (Murzin et al. 1995) rather than being defined based on the directionality of the β-sheets. This could be one more reason why the methods that are based solely on the helix and strand content would have problems with assignments of the α/β and α + β classes. To further investigate this aspect, we generated Table 4 using 25PDB dataset, which shows a side-by-side comparison of the decision tree models generated using just the helix and strand content against the proposed models based on ten features computed from the secondary structure. Although the results are similar for the all-α and all-β classes, we observe that, on average, the proposed methods provide assignments with about 3.5% better overall accuracy. The difference is mostly due to obtaining substantially better results for the α + β class.
- The low accuracy obtained using the assignment method proposed in Chou (1989) can be explained by the large space reserved for irregular proteins. This assignment model assumes the highest threshold values for all three structural classes.
- The best accuracy was achieved by the proposed SSA^{sc} and PSSA^{sc} methods in the case of using the DSSP-assigned and PSI-PRED-predicted secondary structure as the input, respectively. The improvement over the second-best assignment method by Nakashima et al. (1986) and Eisenhaber et al. (1996) is 3.2% in the case of SSA^{sc} and 1.6% for PSSA^{sc}. This improvement translates into 53 and 28 more correct predictions for SSA^{sc} and PSSA^{sc}, respectively. Most importantly, we emphasize that the proposed assignment method is capable of distinguishing the α/β and α + β classes using the 1D secondary structure, which is a significant advantage over the existing secondary structure based assignment methods.

The SSC^{sc} method was applied to classify several proteins for which the existing assignment methods provide incorrect classifications. We included one all-α, one all-β, and

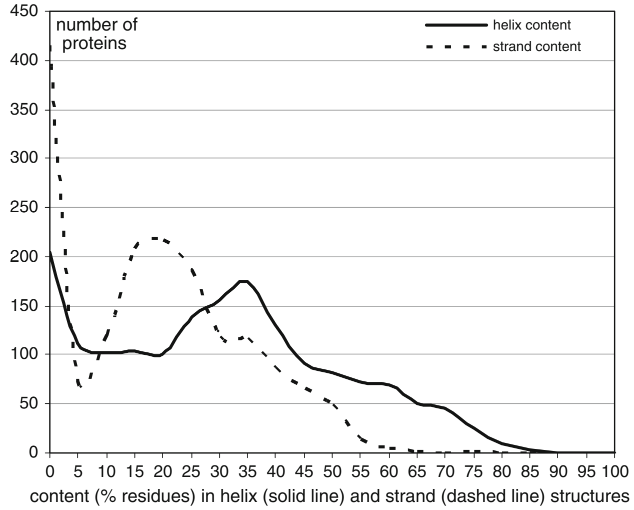


Fig. 3 Distribution of the helix (*solid line*) and strand (*dashed line*) content based on Dictionary of Secondary Structure of Proteins (DSSP)-assigned secondary structure for proteins in the 25PDB dataset

one mixed (α + β) class protein (see Fig. 5). In case of the 1PP7 chain, which was classified in SCOP as all-α protein, the existing methods were misled by the two small β-sheet and categorized this protein either as mixed or irregular class. Similarly, in case of 1R6J protein, the existing method categorized it as mixed or irregular, while in fact in spite of inclusion of a small helix this protein belong to all-β class. Finally, the 1EF5 protein that belongs to α + β class was incorrectly classified as all-β or irregular class. The reason for these mistakes is relatively low amount of helices in this protein. In all three cases, the SSA^{sc} method provided the predictions that agree with the SCOP assignment.

Comparison with existing sequence-based structural class prediction methods

The SSA^{sc} and PSSA^{sc} methods were compared against a set of recently published sequence-based structural class prediction methods, including methods based on a single classification algorithm (Cai et al. 2003; Kedarisetti et al. 2006b; Kurgan and Homaeian 2006; Jahandideh et al. 2007; Kurgan and Chen 2007) and on an ensemble of classifiers (Cai et al. 2006; Dong et al. 2006; Kedarisetti et al. 2006b) (Table 5). Although the SSA^{sc} has the advantage of using the secondary structure assigned with DSSP, which is unavailable for the sequence-based methods, in the case of PSSA^{sc} the secondary structure comes from PSI-PRED, effectively making it a sequence-based prediction/assignment method.

Three important observations can be made:

- The secondary structure-based methods provide more accurate predictions/assignments than the sequence-based methods. For PSSA^{sc}, which also belong to the group of sequence-based methods, the improvement over the sequence-based methods listed in Table 5 equals 12.3%. This improved accuracy is achieved for the prediction of all four structural classes, with the biggest difference being for the α + β class. These improvements can be attributed to the high quality of the predicted secondary structure segments that are used by PSSA^{sc}.
- The proposed sequence representation that includes ten features gives a 2.3 and 2.6% better accuracy in the case of using DSSP-assigned and PSI-PRED-predicted secondary structure, respectively, and when compared against using just the helix and the strand contents (see the last four rows in Table 5). These improvements are

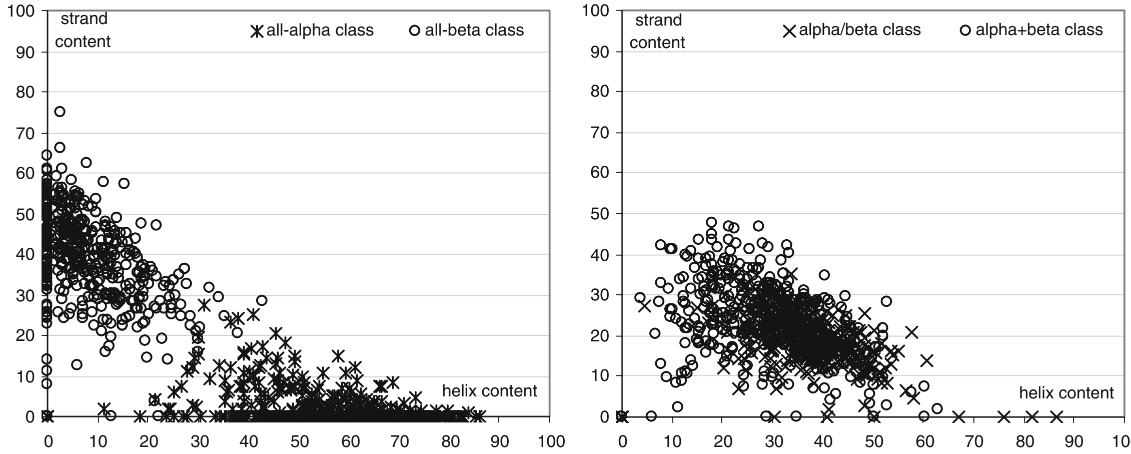


Fig. 4 Scatter plot of helix content (*X*-axis) against strand content (*Y*-axis) based on DSSP-assigned secondary structure for proteins in the 25PDB dataset. The *left panel* shows the plot for the all-α and all-β classes, the *right panel* shows the plots for the α/β and α + β classes

Table 4 Comparison of the accuracy of structural class assignment for the four classes between the proposed SSA^{sc} and PSSA^{sc} methods and methods based solely on the strand and helix content

Assignment method	Source of secondary structure	Accuracy (%)				
		All-α	All-β	α/β	$\alpha + \beta$	Overall
Assignment method using helix and strand content only	DSSP	92.6	84.4	74.1	59.2	77.8
SSA^{sc} using ten features	DSSP	92.6	83.7	65.9	80.5	81.5
Assignment method using helix and strand content only	PSI-PRED	88.7	78.3	75.4	63.0	76.4
PSSA^{sc} using ten features	PSI-PRED	94.6	76.3	73.1	74.4	80.0

due to the use of information on the count, size and location-based content of the strand, helix, and coil structures.

– Most importantly, very similar quality is obtained with both DSSP-assigned and PSI-PRED-predicted secondary structure, i.e., usage of the predicted structure results in only about a 1.5% lower overall accuracy. We note that SSA^{sc} obtains better predictions for all-β and α/β classes, while PSSA^{sc} is better for the $\alpha + \beta$ class. This is most likely the result of the relatively lower quality of predictions of strands, i.e., 66%, by PSI-PRED when compared with the predictions of helices, i.e., 86% (Birzele and Kramer 2006). This result shows that the predicted secondary structure constitutes a high-quality input for the purpose of the structural class assignment.

The predictions shown in Table 5 concern a challenging dataset of sequences with very low sequence identity and, therefore, the reported accuracies are lower than the accuracies reported in some other contributions that include proteins with higher sequence identity (Kurgan and Homaeian 2006). We note that a number of other sequence-based prediction methods were not included in this experimental comparison; they include methods described in Chou and Zhang (1994), Chou (1995), Chou and Zhang (1995), Chou and Maggiora (1998), Chou et al. (1998), Zhou (1998), Cai and Zhou (2000), Wang and Yuan (2000), Cai et al. (2001, 2002a, b), Jin et al. (2003), Feng et al. (2005), Shen et al. (2005), Niu et al. (2006), Xiao et al. (2006a), Cao et al. (2006), Du et al. (2006), Sun and Huang (2006), Zhang and Ding (2007), Chen et al. (2008), and Zhang et al. (2008). Since the proposed method is not meant to compete with the sequence-based predictions, but rather to establish rules that allow the assignment of the class from the secondary structure content, we limited our comparative study to a representative set (which cover a variety of classification methods) of recently published methods.

Conclusion

We propose a novel, accurate method for assigning structural classes based on the 1D secondary structure. We describe two assignment models: SSA^{sc}, which is based on secondary structure assigned from the tertiary structure by DSSP, and PSSA^{sc}, which is based on secondary structure predicted from the sequence by PSI-PRED. The models were designed based on a large set of over 1600 sequences characterized by low pairwise identity. Based on an extensive empirical evaluation and a comparison with seven other secondary structure-based and ten sequence-based structural class assignment/prediction methods, we arrived at several interesting conclusions.

Firstly, the structural class can be successfully assigned/predicted based on 1D-secondary structure, i.e., secondary structure assigned to each residue without the knowledge of its spatial arrangement. Such an assignment is characterized by an accuracy of 76% in the case of the SSA^{sc} and 75% in the case of PSSA^{sc} when compared against the structure-based assignment performed in SCOP. The proposed assignment models are shown to be more accurate than the considered secondary structure-based and sequence-based methods in terms of structural class assignment. We also note that our assignment models, which are based on decision trees, are simple, i.e., small trees that use only ten features, and explicit, i.e., encoded in a human-readable format.

Secondly, the early definitions of structural classes, which were developed before the SCOP was proposed, can be also successfully used to automate the assignment of the classes based on the 1D secondary structure. Although these methods are characterized by a lower accuracy of the assignment when compared with the proposed models, i.e., the differences vary between 3.2 and 52.9% for SSA^{sc} and between 1.6 and 46.7% for PSSA^{sc}, their underlying model is very simple and intuitive. At the same time, we note that these methods are capable only of assigning three types of structural classes—all-α, all-β, and mixed—while the proposed method can also distinguish between the α/β and $\alpha + \beta$ classes.

Lastly, we show that the use of the secondary structure predicted from the sequence (instead of the secondary structure derived from the tertiary structure) does not have a detrimental effect on the quality of structural class assignment. The corresponding difference in assignment

Fig. 5 Ribbon structures of the 1PP7, 1R6J, and 1EF5 domains together with their structural class assignments computed by existing methods and the proposed SSA^{sc} method. The structures were prepared with MBT package (Moreland et al. 2005)

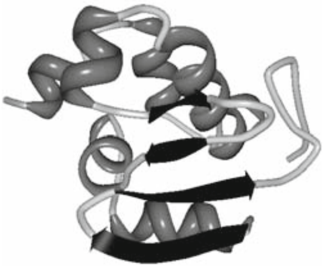
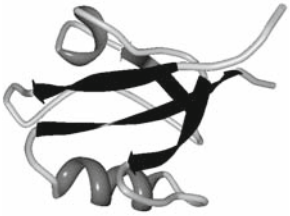
		assignment method	**assigned class**
1PP7 chain U		(Nakashima et al., 1986) and (Eisenhaber et al., 1996)	mixed
		(Klein and DeLisi, 1986)	mixed
		(Chou, 1989)	irregular
		(Kneller et al., 1990)	mixed
		(Chou, 1995)	mixed
		(Liu and Chou, 1998)	mixed
		SSA^{sc}	all-α
		actual structural class in SCOP	all-α
1R6J chain A		(Nakashima et al., 1986) and (Eisenhaber et al., 1996)	mixed
		(Klein and DeLisi, 1986)	mixed
		(Chou, 1989)	irregular
		(Kneller et al., 1990)	mixed
		(Chou, 1995)	mixed
		(Liu and Chou, 1998)	mixed
		SSA^{sc}	all-β
		actual structural class in SCOP	all-β
1EF5 chain A		(Nakashima et al., 1986) and (Eisenhaber et al., 1996)	all-β
		(Klein and DeLisi, 1986)	irregular
		(Chou, 1989)	irregular
		(Kneller et al., 1990)	all-β
		(Chou, 1995)	irregular
		(Liu and Chou, 1998)	irregular
		SSA^{sc}	α+β (mixed)
		actual structural class in SCOP	α+β

Table 5 Comparison of the accuracy of structural class prediction for the four classes between the proposed SSA^{sc} and PSSA^{sc} methods and recently proposed sequence based methods

Algorithm	Reference	Accuracy (%)				
		All-α	All-β	α/β	α + β	Overall
SVM (Gaussian kernel)	Cai et al. (2003)	67.9	59.1	58.1	27.7	53.0
LogicBoost with decision tree	Cai et al. (2006)	51.9	53.7	46.5	32.4	46.1
LogitBoost with decision stump	Dong et al. (2006)	63.2	53.5	50.9	32.4	50.0
Multinomial logistic regression	Jahandideh et al. (2007)	56.9	44.2	42.2	17.7	40.2
Multinomial logistic regression	Kedarisetti et al. (2006b)	69.9	65.3	66.5	38.4	60.0
SVM (RBF kernel)		70.2	61.6	67.6	39.6	59.8
StackingC ensemble		73.4	67.3	69.1	29.8	59.9
Multinomial logistic regression	Kurgan and Homaeian (2006)	69.1	60.5	59.5	38.1	56.7
SVM (1st order polyn. kernel)	Kurgan and Chen (2007)	77.7	66.8	60.7	45.4	62.8
Linear logistic regression		74.7	66.4	62.7	45.8	62.4
Decision tree based only on helix and strand content (DSSP-based assignment)	This paper	90.3	81.9	67.6	55.3	74.1
Decision tree based only on helix and strand content (PSI-PRED-based assignment)	This paper	88.3	76.3	72.0	53.3	72.5
SSA^{sc} (DSSP-based assignment)	This paper	90.7	80.1	68.8	64.1	76.4
PSSA^{sc} (PSI-PRED-based assignment)	This paper	89.6	74.7	65.3	68.5	75.1

accuracy equals 1.5%. Therefore, such a sequence-based assignment model constitutes a feasible alternative to automate the assignment of structural classes without any knowledge of the protein structure.

Acknowledgments This research was funded in part by NSERC Canada. TZ and HZ would like to acknowledge financial support provided by National Education Committee of China. S. Shen and J Ruan were supported by NSFC (grant no. 10671100), Liuhui Center for applied mathematics, and the joint program of Tianjin and Nankai Universities.

References

Andreeva A, Howorth D, Brenner S, Hubbard T, Chothia C, Murzin A (2004) SCOP Database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32:D226–D229

Bahar I, Atilgan AR, Jernigan RL, Erman B (1997) Understanding the recognition of protein structural classes by amino acid composition. Proteins 29:172–185

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

Birzele F, Kramer S (2006) A new representation for protein secondary structure prediction based on frequent patterns. Bioinformatics 22:2628–34

Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT (2005) Protein structure prediction servers at university college London. Nucleic Acids Res 33:W36–38

Cai YD, Zhou GP (2000) Prediction of protein structural classes by neural network. Biochimie 82:783–85

Cai YD, Liu XJ, Xu X, Zhou GP (2001) Support vector machines for predicting protein structural class. BMC Bioinformatics 2:3

Cai YD, Liu XJ, Xu XB, Chou KC (2002a) Prediction of protein structural classes by support vector machines. Comput Chem 26:293–296

Cai YD, Hu J, Liu XJ, Chou KC (2002b) Prediction of protein structural classes by neural network method. J Mol Des 1:332–338

Cai YD, Liu XJ, Xu XB, Chou KC (2003) Support vector machines for prediction of protein domain structural class. J Theor Biol 221:115–20

Cai YD, Feng KY, Lu WC, Chou KC (2006) Using logitboost classifier to predict protein structural classes. J Theor Biol 238:172–6

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

Carlacci L, Chou KC, Maggiora GM (1991) A heuristic approach to predicting the tertiary structure of bovine somatotropin. Biochemistry 30:4389–4398

Cedano J, Aloy P, P'erez-Pons JA, Querol E (1997) Relation between amino acid composition and cellular location of proteins. J Mol Biol 266:594–600

Chen K, Kurgan L (2007) PFRES: protein fold classification by using evolutionary information and predicted secondary structure. Bioinformatics 23:2843–2850

Chen K, Kurgan L, Ruan J (2008) Prediction of protein structural class using novel evolutionary collocation-based sequence representation. J Comput Chem. doi: 10.1002/jcc.20918

Chou KC (1992) Energy-optimized structure of antifreeze protein and its binding mechanism. J Mol Biol 223:509–517

Chou KC (1995) A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. Proteins 21:319–344

Chou KC (2005a) Prediction of G-protein-coupled receptor classes. J Proteome Res 4:1413–1418

Chou KC (2005b) Progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr Protein Pept Sci 6:423–436

Chou KC, Cai YD (2004) Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun 321:1007–1009

Chou KC, Elrod DW (1999) Protein subcellular location prediction. Protein Eng 12:107–118

Chou KC, Maggiora GM (1998) Domain structural class prediction. Protein Eng 11:523–538

Chou KC, Shen HB (2007a) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360:339–345

Chou KC, Shen HB (2007b) Recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16

Chou KC, Shen HB (2007c) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Commun 357:633–640

Chou KC, Shen HB (2008) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Protoc3:153–162

Chou KC, Zhang CT (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 269:22014–20

Chou KC, Zhang CT (1995) Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349

Chou KC, Liu W, Maggiora GM, Zhang CT (1998) Prediction and classification of domain structural classes. Proteins 31:97–103

Chou JJ, Zhang CT (1993) A joint prediction of the folding types of 1,490 human proteins from their genetic codons. J Theor Biol 161:251–262

Chou PY (1989) Prediction of protein structural classes from amino acid composition. In: Fasman GD (ed) Prediction of protein structure. Plenum Press, New York, pp 549–586

Dong L, Yuan Y, Cai T (2006) Using bagging classifier to predict protein domain structural class. J Biomol Struct Dyn 24:239–42

Du QS, Jiang ZQ, He WZ, Li DP, Chou KC (2006) Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. J Biomol Struct Dyn 23:635–640

Eisenhaber F, Frömmel C, Argos P (1996) Prediction of secondary structural content of proteins from their amino acid composition alone. II The paradox with secondary structural class. Proteins 25:169–179

Feng KY, Cai YD, Chou KC (2005) Boosting classifier for predicting protein domain structural class. Biochem Biophys Res Commun 334:213–7

Fuchs PF, Alix AJ (2005) High accuracy prediction of beta-turns and their types using propensities and multiple alignments. Proteins 59:828–39

Garg A, Kaur H, Raghava GP (2005) Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins 61:318–24

Gromiha MM (2005a) Motifs in outer membrane protein sequences: applications for discrimination. Biophys Chem 117(1):65–71

Gromiha MM (2005b) A statistical model for predicting protein folding rates from amino acid sequence with structural class information. J Chem Inf Model 45(2):494–501

Gromiha M, Selvaraj S (1998) Protein secondary structure prediction in different structural classes. Protein Eng 11:249–251

Gromiha MM, Suwa M (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy. Bioinformatics 21:961–8

Gromiha MM, Selvaraj S, Thangakani AM (2006) A statistical method for predicting protein unfolding rates from amino acid sequence. J Chem Inf Model 46:1503–1508

He H, McAllister G, Smith TF (2002) Triage protein fold prediction. Proteins 48:654–63

Hobohm U, Sander C (1994) Enlarged representative set of protein structures. Protein Sci 3:522

Ivankov DN, Finkelstein AV (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Natl Acad Sci USA 101:8942–4

Jahandideh S, Abdolmaleki P, Jahandideh M, Sadat Hayatshahi SH (2007) Novel hybrid method for the evaluation of parameters contributing in determination of protein structural classes. J Theor Biol 244:275–81

Jin L, Fang W, Tang H (2003) Prediction of protein structural classes by a new measure of information discrepancy. Comput Biol Chem 27:373–80

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:95–202

Kabsch W, Sander C (1983) Dictionary of protein secondary structures: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637

Kedarisetti KD, Kurgan L, Dick S (2006a) A comment on 'prediction of protein structural classes by a new measure of information discrepancy'. Comput Biol Chem 30:393–4

Kedarisetti KD, Kurgan L, Dick S (2006b) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348:981–8

Klein P, DeLisi C (1986) Prediction of protein structural class from the amino acid sequence. Biopolymers 25:1659–1672

Kneller DG, Cohen FE, Langridge R (1990) Improvements in secondary structure prediction by enhanced neural networks. J Mol Biol 214:171–182

Kurgan L, Homaeian L (2006) Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recognit 39:2323–43

Kurgan L, Chen K (2007) Prediction of protein structural class for the twilight zone sequences. Biochem Biophys Res Commun 357:453–60

Kuznetsov IB, Gou Z, Li R, Hwang S (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins 64:19–27

Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261:552–557

Lin K, Simossis V, Taylor W, Heringa J (2005) A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics 21:152–9

Liu W, Chou KC (1998) Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. Protein Chem 17:209–217

Martin J, Letellier G, Marin A, Taly J-F, de Brevern AG, Gibrat J-F (2005) Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct Biol 5:17

Moreland JL, Gramada A, Buzko OV, Zhang Q, Bourne PE (2005) The molecular biology toolkit (MBT): a modular platform for developing molecular visualization applications. BMC Bioinformatics 6:21

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

Nakashima H, Nishikawa K, Ooi T (1986) The folding type of a protein is relevant to the amino acid composition. J Biochem 99:153–162

Niu B, Cai YD, Lu WC, Zheng GY, Chou KC (2006) Predicting protein structural class with AdaBoost learner. Protein Pept Lett 13:489–492

Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Francisco

Shen HB, Yang J, Liu X-J, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 334:577–81

Shen HB, Chou KC (2007a) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun 364:53–59

Shen HB, Chou KC (2007b) Signal-3L: a 3-layer approach for predicting signal peptide. Biochem Biophys Res Comm 363:297–303

Shen HB, Chou KC (2007c) Using ensemble classifier to identify membrane protein types. Amino Acids 32:483–488

Song J, Burrage K (2006) Predicting residue-wise contact orders in proteins by support vector regression. BMC Bioinformatics 7:425

Song J, Yuan Z, Tan H, Huber T, Burrage K (2007) Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. Bioinformatics 23:3147–54

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30:469–475

Wang Y, Xue Z, Xu J (2006) Better prediction of the location of alpha-turns in proteins with support vector machine. Proteins 65:49–54

Wang Z-X, Yuan Z (2000) How good is the prediction of protein structural class by the component-coupled method? Proteins 38:165–175

Wen Z, Li M, Li Y, Guo Y, Wang K (2007) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32:277–283

Witten IH, Frank E (2005) Data mining. Practical machine learning tools and techniques. 2nd edn. Morgan Kaufmann, San Francisco

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28:57–61

Xiao X, Shao S, Huang Z, Chou KC (2006a) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comp Chem 27:478–82

Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC (2006b) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30:49–54

Zhang CT, Zhang Z, He Z (1998) Prediction of the secondary structure contents of globular proteins based on three structural classes. J Protein Chem 17:261–72

Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids 33:623–629

Zhang TL, Ding YS, Chou KC (2008) Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern. J Theor Biol 250:186–193

Zhang Z, Sun ZR, Zhang CT (2001) A new approach to predict the helix/strand content of globular proteins. J Theor Biol 208:65–78

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17:729–38