Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
BBRC
Biochemical and Biophysical Research Communications 348 (2006) 981–988
www.elsevier.com/locate/ybbrc

Classifier ensembles for protein structural class prediction with varying homology

Kanaka Durga Kedarisetti, Lukasz Kurgan *, Scott Dick

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

Received 18 July 2006

Abstract

Structural class characterizes the overall folding type of a protein or its domain. A number of computational methods have been proposed to predict structural class based on primary sequences; however, the accuracy of these methods is strongly affected by sequence homology. This paper proposes, an ensemble classification method and a compact feature-based sequence representation. This method improves prediction accuracy for the four main structural classes compared to competing methods, and provides highly accurate predictions for sequences of widely varying homologies. The experimental evaluation of the proposed method shows superior results across sequences that are characterized by entire homology spectrum, ranging from 25% to 90% homology. The error rates were reduced by over 20% when compared with using individual prediction methods and most commonly used composition vector representation of protein sequences. Comparisons with competing methods on three large benchmark datasets consistently show the superiority of the proposed method.

Keywords: SCOP structural class; Sequence homology; Protein primary sequence; Feature selection; Classifier ensemble

Knowledge of protein structure plays a key role in analysis of protein functions, protein binding, rational drug design, and many other related fields and applications. The structure is organized on several levels, which include secondary, tertiary, and quaternary structure. Despite assuming a complex three-dimensional structure and having a relatively irregular surface, proteins are characterized by quite simple and regular local folding patterns even if they bear different biological functions. Protein structural class prediction is based on identifying these simple patterns, and applying them to proteins with unknown structures.

In view of this, Levitt and Chothia [36] introduced the concept of protein structural class about three decades ago based on a visual inspection of polypeptide chain topologies in a dataset of 31 globular proteins and divided these proteins into four structural classes: all-α, all-β, α/β, and α + β. The all-α and all-β classes are formed by sequences that consist of virtually only α-helices and β-strands, respectively. The α/β class includes proteins with both α-helices and β-strands, where the latter mainly include parallel β-sheets. Finally, the α + β class includes proteins with α-helices and antiparallel β-sheets. One of the most accurate classifications of protein structural classes can be found in the curated SCOP (structural classification of proteins) database [39]. The basic structural unit of classification in the SCOP database is either the entire sequence or a protein domain (structurally conserved fragment of the sequence), depending on if the domains that compose the sequence belong to the same or different structural classes.

Over the last twenty years, the number of known protein structures has significantly increased. As a result, numerous computational methods have been developed to predict protein structural class based on the primary amino acid sequence, beginning in the 1980s [29,40], some

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

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.07.141

advancements made in 1990s [12,19,20,38,48], and the most recent methods [8,9,16,22,26,43,46]. The early methods were tested on very limited protein sets, which resulted in very low performance. More recently, performance results have ranged from relatively low to high. These methods were often tested on small datasets, characterized by different homology between sequences, and were not reliably compared with other methods on common data [31]. In addition, current protein structural class prediction results are quite poor compared to other protein secondary structure prediction methods. The goal of our research is to rectify these shortcomings, by applying an ensemble of classification algorithms and developing a compact feature representation of protein sequences. We proceed in three stages:

1. Design of a sequence representation via feature selection, and selection and setup of an optimal classifier ensemble.
2. Validation of the proposed method. We quantify the added value of the feature representation and of the ensemble. We evaluate the method for high- and low-homology datasets.
3. Comparison of the proposed method with other state-of-the-art prediction methods on three large benchmark datasets characterized by different sequence homologies.

The validation and comparison were performed using three commonly used test procedures: out-of-sample 10-fold cross-validation; *n*-fold cross validation (jackknife test); and in-sample resubstitution. The latter test was reported only for consistency with prior work; in-sample errors should not be used to evaluate prediction models [31]. The 10-fold cross-validation was applied to design the prediction method, while the jackknife test was used to perform comparison with the competing methods.

Materials and methods

Structural classes

The original protein structural classification scheme defined by Levitt and Chothia included four classes. This scheme has been modified multiple times by changing the thresholds for the amount of helices and strands that define a structural class. In 1986 Nakashima and colleagues defined five structural classes [40], which were redefined with different thresholds by Chou in 1995 [13]. Chou's criticism of Nakashima's classification was that the thresholds for all-α proteins and all-β proteins were not large enough to reflect these two structural classes. Chou also used dictionary of secondary structure of proteins (DSSP) [27] to define content of the secondary structures. Eisenhaber and colleagues proposed another definition in 1996 [21], which merges the $\alpha + \beta$ and the α/β classes into a "mixed" class. These classifications omit irregular or "ξ" proteins, which are small in number.

The SCOP classification is performed manually, using structural elements located in individual domains within the protein. Researchers claim that the SCOP classification is more natural and provides more reliable information to study protein structural classes when compared to the above classifications [15,39,46]. The SCOP classification currently includes eleven classes [2]: (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; and (11) designed proteins. Our research focuses on computational prediction of the first four categories, as they include the great majority of protein sequences and were the basis for most comparable approaches [7,9,26,31,43,44,46].

Related work

Structural class prediction is usually performed in two steps. First, the primary AA sequence is transformed into a fixed-length feature vector. Next, the feature vectors are fed to a classification algorithm to perform the prediction. The early computational prediction methods represented the primary sequence using only the composition vector- and threshold-based class definitions, and applied discriminant analysis with simple distance definitions as the classification algorithm. Examples include the Euclidean distance [40], the Hamming distance [29], and the Mahalanobis distance [12]. Later prediction methods used more complex classification algorithms, and the same composition vector-based representation. Examples include the maximum component coefficient principle algorithm [48], least correlation angle algorithm [11], fuzzy clustering [43,49], artificial neural networks [19,20,38], vector decomposition [21], the component-coupled geometric classification algorithm [15], Bayesian classification [46], and most recently support vector machines [9] and logit boosting [22]. The most noticeable progress among these algorithms is the inclusion of the coupling effect among different AAs [14,15]. Recent works also improve structural class prediction by using alternative sequence representations. Examples include auto-correlation functions based on non-bonded residue energy [7], polypeptide composition [26,37,44], and functional domain composition [16]. However, these algorithms are often only tested on very small datasets, with uncontrolled (often high) sequence homology. This tends to result in an overestimated prediction accuracy. In addition, they do not perform reliable comparison with other algorithms on common datasets and some incorrectly perform out-of-sample tests [28,31]. Finally, they use a relatively limited feature-based sequence representation.

Motivation

There are several factors that motivate development of a new method for protein structural class prediction:

- The accuracy of secondary structure prediction [23] can be significantly improved by incorporating knowledge of structural classes. This can lead to a reduction of the search space of possible conformations of the tertiary structure [3,14].
- The huge and growing gap between known protein sequences (over 2 million protein sequences stored in National Center for Biotechnology Information (NCBI) database) and known protein structures (37658 structures stored in Protein Data Bank (PDB) [4] as on July 11, 2006). Development of a reliable computational method for prediction of structural classes for new, unclassified protein sequences is essential.
- Sequence homology has a significant impact on prediction accuracy. While this is well known, this factor is generally uncontrolled in prediction experiments. For example, the commonly used "359" dataset [15] is highly homologous, leading to over 80% prediction accuracy [7,9]. However, low accuracies (often about 50%) are the norm for low-homology datasets [31,46]. In addition, some researchers assert that prediction methods are only viable for datasets with less than 30% homology [46], as sequence-alignment-based prediction is effective for higher-homology datasets.
- Several empirical studies have shown that the performance of ensemble machine learning approaches is superior to individual learning algorithms in various applications in structural biology including protein fold classification [10,18,45]. To the best of our knowledge, ensembles have never been applied for structural class prediction. A multi-classi-

fier method for structural class prediction was developed by Feng et al. [22], but they did not consider heterogeneous classifiers (a key strength of ensemble methods).

- Finally, existing methods for structural class prediction have relatively low accuracy when compared to other secondary structure prediction methods.

We propose a novel ensemble prediction method that is more accurate than competing individual classifiers, and which provides good performance over a range of sequence homologies.

Overview of the proposed method

An overview of the proposed method is shown in Fig. 1. We begin by creating an extensive feature-based sequence representation (including the composition vector, autocorrelations of hydrophobicity indices, etc.) and then reducing the dimensionality of the feature space via feature selection. The resulting features vectors are fed to four heterogeneous classifiers, which each output a predicted structural class. Finally, these predictions are combined by a specialized "voting" module that outputs the final prediction.

Datasets

Source datasets. We consider a total of six datasets characterized by different sequence homology in this study. Out of these, four datasets with strictly controlled sequence homology, i.e., 25PDB, 50PDB, 70PDB, and 90PDB with the sequence homology of 25%, 50%, 70%, and 90%, respectively, are used for the design of the prediction method. The 25PDB dataset [25] and the remaining two datasets were used in previous studies, and are used to compare the designed method with other competing methods. These two datasets are "359" [15] and "1189" [46]. The six datasets were processed using the PDB release as of June 2005 and were filtered based on the four SCOP classes, i.e., sequences belonging to other or undetermined SCOP labels were removed. The class distribution within these datasets is given in Table 1. Next, we briefly describe how these datasets were generated.

The "359" dataset includes highly homologous sequences (over 95% homology) and was extensively used to evaluate various structural class prediction methods. "1189" and 25PDB datasets are both low-homology datasets (40% and 25% homology, respectively), which have been used in previous studies of structural class prediction methods. The other three datasets are similar to 25PDB in that homology in these datasets is strictly controlled. They are based on the clusters50, clusters70, and clusters90 lists published on the PDB web site as of June 2005. The latter datasets were extracted using the CD-hit (Cluster Database at High Identity with Tolerance) program [32]. After removing duplicate protein domains and filtering based on the four SCOP classes, these three datasets include 3642, 4243, and 5212 sequences, respectively.

Table 1
SCOP structural class distribution in the datasets used in this study

Dataset	All-α	All-β	α/β	α + β	Total
359	76	81	94	81	332
1189	223	294	334	241	1092
25PDB	443	443	346	441	1673
50PDB	637	921	1214	870	3642
70PDB	742	1101	1389	1318	4243
90PDB	840	1657	1553	1162	5212

Preparation of the controlled homology datasets. The 25PDB, 50PDB, 70PDB, and 90PDB datasets were constructed to explicitly control dataset homology in our study. However, in addition to controlling homology, we had to control other characteristics of the datasets, especially *class distribution*. All machine learning algorithms are negatively impacted by imbalance between different classes; in general, the more prevalent classes will tend to dominate the learning process. The original 25PDB, 50PDB, 70PDB, and 90PDB datasets are highly unbalanced with respect to the distribution of the SCOP labels. Additionally, the original 50PDB, 70PDB, and 90PDB datasets contain a significant amount of overlapping sequences, i.e., protein domains from clusters50 are listed in clusters70 and protein domains from clusters70 are included in cluster90. These characteristics make the comparison of prediction accuracies between the datasets very difficult. Therefore, these datasets were transformed into balanced datasets with minimal overlaps (25HPDB, 50HPDB, 75HPDB, and 90HPDB), in three steps:

1. We filtered all 50PDB domains from 70PDB to obtain the 70–50PDB dataset. Similarly, we filtered all 70PDB domains from 90PDB and obtained the 90–70PDB dataset.
2. Number of sequences for the structural class that has minimum number of records for each of the 25PDB, 50PDB, 70PDB, and 90PDB datasets was found.
3. The number of sequences for each SCOP class in a dataset was equalized.
   (a) Since, there is no overlap between 25PDB and 50PDB, the corresponding balanced datasets, i.e., 25HPDB and 50HPDB, were created by selecting all sequences from the class with the minimum number of records and randomly (without replacement) drawing the same number of sequences for the remaining classes.
   (b) To reduce the overlap, the remaining two balanced datasets, i.e., 70HPDB and 90HPDB, were created by adding sequences from 50PDB to the 70–50PDB dataset, and from 70PDB to the 90–70PDB dataset. For the class with the minimum number of records all sequences were included, and for the remaining classes the added sequences were randomly drawn (without replacement) from 50PDB to 70PDB, respectively.

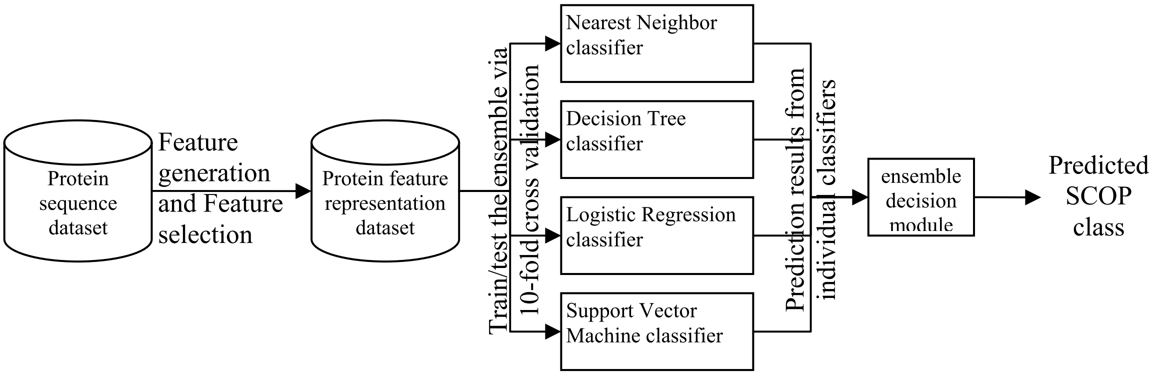
Protein sequence dataset
Feature generation and Feature selection
Protein feature representation dataset
Train/test the ensemble via 10-fold cross validation
Nearest Neighbor classifier
Decision Tree classifier
Logistic Regression classifier
Support Vector Machine classifier
Prediction results from individual classifiers
ensemble decision module
Predicted SCOP class

Fig. 1. Proposed ensemble method for the structural class prediction.

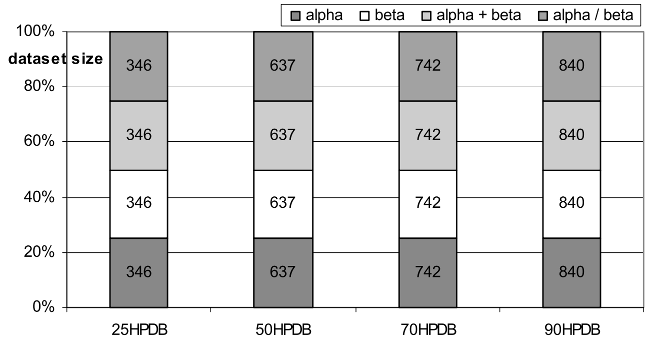

Fig. 2. Distribution of class labels in the 25HPDB, 50HPDB, 70HPDB, and 90HPDB datasets.

As a result, 1384, 2548, 2968, and 3360 domains were included in the 25HPDB, 50HPDB, 70HPDB, and 90HPDB datasets, respectively, see Fig. 2.

Feature representation of protein sequences. The protein sequences were initially represented using a comprehensive set of 122 features, grouped into 20 feature sets (see Table 2). This was reduced to 34 features through the application of several feature selection algorithms on the original set of 122 features.

Our aim in feature selection was to develop a compact and accurate representation of protein sequences. The 34 features used in our final representation were selected using the "1189" dataset and three feature selection methods: (1) feature subset consistency (FSC) method selects a subset of features for maximum consistency in the class values when the training instances are projected onto the feature subset [34]. (2) Wrapper subset selection (WSS) method is a classification-based wrapper that uses the Naïve Bayes algorithm [30]. (3) Feature correlation (FC) method selects a subset of features that are highly correlated with the class while having low inter-correlation [24]. Each feature selection method was executed using 10-fold cross-validation test. Individual "best" features were selected in two steps:

1. Selection of best feature sets (see Table 3). First, we compute the number of features for each feature set that are selected in at least 5-folds by any of the three selected feature selection methods. These features are strongly correlated with the SCOP labels. Next, to select only significant properties (feature sets), every feature set for which at least 50% of the features were selected by this rule is found (bolded and underlined in Table 3).

Table 2
Feature representation

Feature set ID	Feature set name	Abbr.	# features	Brief description
1	Sequence length	L	1	Number of AAs in a protein sequence
2	Molecular weight	M	1	Sum of the molecular weights of the neutral, free AAs
3	Isoelectric point	I	1	Sum of isoelectric point values of the AAs
4	Composition vector	CV	20	normalized composition percentage of each AA
5	1st order composition moment vector	1-CMV	20	1st order composition vector that incorporates composition and position of each AA
6	2nd order composition moment vector	2-CMV	20	2nd order composition vector that incorporates composition and position of each AA
7	R group	RG	5	Combines hydropathy, molecular weight, and p*I*
8	Exchange group	EXG	3	Some AAs can be substituted by other without impact on the structure
9	Hydrophobic group	HG	2	Divides AAs into hydrophobic and hydrophilic
10	Electronic group	EG	5	Divides AAs into neutrals, electron donors or acceptors
11	Chemical group	C	19	Chemical groups associated with the AAs
12	Other groups	O	7	Divides AAs into charged, polar, aromatic, small, tiny, bulky, and polar uncharged
13	Auto-correlation-1	AC1	6	Autocorrelations computed using the Black and Mould hydrophobic index [5]
14	Sum of hydrophobicity-1	SH1	1	Sum of Black and Mould hydrophobicity index
15	Average of hydrophobicity-1	AH1	1	Average of Black and Mould hydrophobicity index
16	Sum of the running three-average of hydrophobicity-1	A3H1	1	Sum of running three-average of Black and Mould hydrophobicity index
17	Auto-correlation-2	AC2	6	Autocorrelations computed using the Eisenberg's hydrophobic index [17]
18	Sum of hydrophobicity-2	SH2	1	Sum of Eisenberg's hydrophobic index
19	Average of hydrophobicity-2	AH2	1	Average of Eisenberg's hydrophobicity index
20	Sum of the running three-average of hydrophobicity-2	A3H2	1	Sum of running three-average of Eisenberg's hydrophobicity index

Table 3
Results of the selection of the best feature sets

	Feature set ID (see Table 2)																				Total # features
	1	2	3	**4**	5	6	**7**	8	**9**	**10**	11	12	**13**	**14**	15	**16**	**17**	**18**	19	**20**	
# features in a set	1	1	1	20	20	20	5	3	2	5	19	7	6	1	1	1	6	1	1	1	122
# features selected by	1	0	0	13	4	3	3	0	1	3	1	0	4	1	0	1	5	1	0	1	42

Any of 3 feature selection methods in at least 5-folds.

Table 4
Results of the selection of the feature-based sequence representation

Feature set abbr.	Total # features	# selected features	Selected features
L	1	1	Sequence length
CV	20	13	Composition of **A**, C, **D**, **E**, **G**, **I**, **L**, **M**, **P**, Q, T, **V**, **Y** residues
RG	5	3	**Non-polar** (A,V,L,I,M,G), positively charged (K,H,R), and **negatively charged** (D,E) residue sets
HG	2	1	Hydrophilic polar with uncharged side chain
EG	5	3	**Electron donor** (D,E,P,A), **weak electron donor** (V,L,I), and neutral (G,H,W,S) residue sets
AC1	6	4	First 4 autocorrelations computed using the Black and Mould hydrophobic index
SH1	1	1	See Table 2
A3H1	1	1	See Table 2
AC2	6	5	First 4 and 6th autocorrelations computed using the Eisenberg's hydrophobic index
SH2	1	1	See Table 2
A3H2	1	1	See Table 2
Total	49	34	Total

2. Selection of individual best features from the best feature sets (see Table 4). To reduce the number of features we select only those features that were selected by the rule in step 1 *and* belong to the best feature sets. This finally results in 34 features shown in Table 4.

In Table 4 we use boldface to indicate relationships between the 34 selected features. The AAs selected from the composition vector mostly belong to the non-polar and negatively charged groups (R group feature set), and to the electron donor groups (electronic group). This shows that polarity, charge, and electron donor properties may have an impact on the folding into a specific structural class. We also observe that hydrophobicity is strongly related to prediction of structural classes. More specifically, a number of autocorrelations, which focus on local conformations and are likely to be sensitive to helical structures, were included in the final 34 features. This agrees with the findings of other researchers with respect to the prediction of protein secondary structure content [33,50].

Ensemble of classifiers. The proposed prediction method consists of four base classifiers that are combined into an ensemble, all of which is implemented in the WEKA environment [47]. We first optimize the parameters for the individual base classifiers for the "1189" dataset using 10-fold cross validation. Two of these classifiers work relatively well with low homology datasets: logistic regression (LgRe), which implements multinomial regression [35], and support vector machines (SVM), which is a kernel-based method [41]. The other two classifiers perform best on high-homology datasets: instance-based (IB1) classifier, which implements the nearest-neighbor algorithm [1], and random forest (RF), which constructs a collection of decision trees [6]. The performance of the four base classifiers using 10-fold cross-validation on the four controlled homology datasets is shown in Fig. 3.

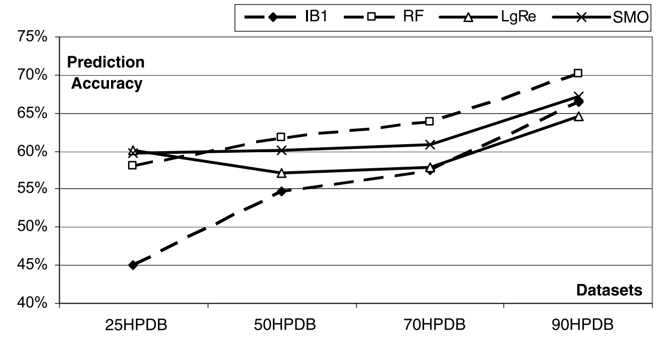

Fig. 3. Accuracy of the four base classifiers on the controlled homology datasets.

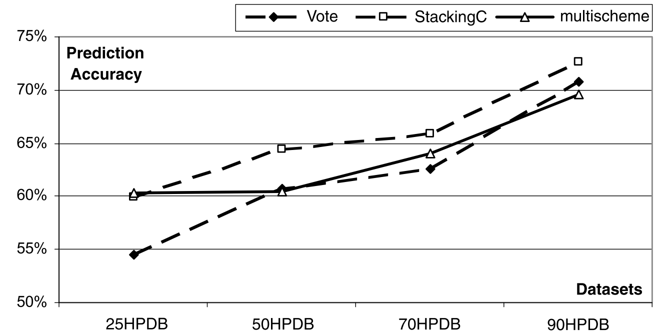

Fig. 4. Accuracy of the three ensembles of classifiers on the controlled homology datasets.

We examined three different ensemble techniques (voting, multi-scheme, and stacking) to select the optimal method to combine the base classifiers. Voting is a generic method that combines classifier results using an unweighted average of the classification probability estimates. Multi-scheme is another generic method that selects one classifier amongst the ensemble to perform the classification based on their performance on the cross-validation test. Finally, stacking uses the predictions of the base classifiers as attributes in a new training set that keeps the class labels. This new meta-training set is used to train a meta classifier that is used to predict the class. We used a recent implementation of stacking, called stackingC, which is characterized by improved performance and superior results for multi-class data [42]. The prediction accuracy of the three ensembles for the controlled homology datasets is compared in Fig. 4.

The results show superiority of the stackingC method. A paired *t*-test between the stackingC and the other two ensemble methods gave *t* values of 5.12 and 2.36 for the voting and multi-scheme methods, respectively. The differences are thus statistically significant at 99% and 95% significance levels for the voting and multi-scheme methods, respectively. Therefore, the stackingC method was used to implement the ensemble.

Results and discussion

Validation of the proposed prediction method

Our aim in this section is to evaluate the added value of the proposed feature representation, as well as that of the ensemble method. We empirically compare our proposed method with the best results of the individual base classifiers, and with the use of the composition vector representation (the most common representation in structural class prediction). We used the controlled-homology datasets and 10-fold cross-validation in all of our experiments, in which we consider four configurations:

1. The proposed ensemble that uses 34 features to represent protein sequences (stackingC-34f).
2. The proposed ensemble that uses CV to represent protein sequences (stackingC-CV).
3. The best results for each of the datasets generated by base classifiers using the 34 feature representation (best base-34f).
4. The best results for each of the datasets generated by base classifiers using CV representation (best base-CV).

Our results are shown in Fig. 5.

The classification accuracies range between 49% and 73%. The *t* value between stackingC with the proposed representation and the *best base-34f*, *stackingC-CV*, and *best base-CV* is 2.73, 12.0, and 24.4, respectively. Therefore, the differences are statistically significant at 95%, 99.9%, and 99.95% levels, respectively. Overall the proposed model improved the classification accuracy by up to 10.8% when compared to using single classifiers and CV representation. This graph yields three main findings:

1. Prediction of structural classes is more difficult on low-homology sequences than for higher-homology data, i.e., the best results for 25HPDB are about 60%, while the best results for 90HPDB are about 73%.
2. The proposed feature-based representation of protein sequences on average reduced the error rates by 20.1% when using the ensemble (on average 8.6% reduction of the average error rate of the *stackingC-CV* method that equals 42.9%) and by 17.9% when using individual classifiers (on average 8.0% reduction of the average error rate of the *best base-CV* method that equals 44.1%). The reductions for individual datasets range between 15.6% and 22.5%.
3. Similarly, the application of the ensemble reduced the errors on average by 2.6% (on average 1.1% reduction of the average error rate of the *best base-CV* method that equals 44.1%) and by 5.3% (on average 1.8% reduction of above the average error rate of the *best base-34f* method that equals 36.1%) when the CV and the proposed representations were used, respectively. The reductions range between −1.5% and 8.7%.

This shows that the biggest reduction of error rates (improvement of accuracy) was achieved through the custom protein sequence representation, but at the same time, classifier ensembles also improved our predictions, especially for sequences characterized by higher homology.

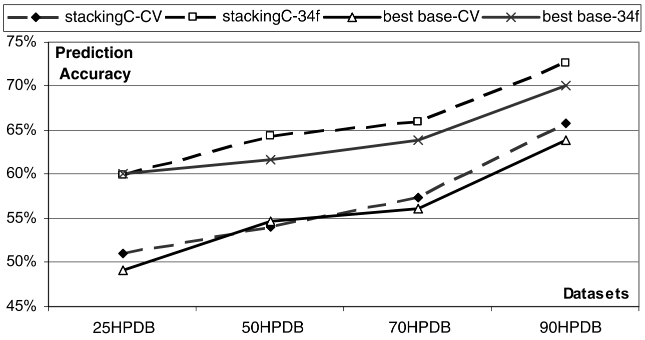

Fig. 5. Evaluation of the proposed protein sequence representation and the ensemble of classifiers.

Table 5
Comparison with other prediction methods (the highlights show results of the proposed method)

Dataset	Classification algorithm	Sequence representation	Accuracy [%]		Reference
			Resubst.	Jackknife	
359	Support vector machine	66 features	100	97.0	Kurgan and Homaeian [31]
	Instance-based classifier	66 features	100	97.0	Kurgan and Homaeian [31]
	StackingC ensemble	34 features	100	96.4	This paper
	Information discrepancy classifier	Polypeptides composition vector	100	95.8	Jin et al. [26]
	Support vector machine	CV	93.0	95.2	Cai et al. [9]
	Component Coupled	Auto-correlations	96.7	90.5	Bu et al. [7]
	Geometric classifier	CV	94.4	84.7	Bu et al. [7]
	Geometric classifier	CV	94.3	84.1	Chou and Maggiora [15]
1189	StackingC ensemble	34 features	95.2	58.9	This paper
	Logistic regression	66 features	62.0	53.9	Kurgan and Homaeian [31]
	Bayes classifier	CV	63.8	53.8	Wang and Yuan [46]
	Support vector machine	CV	57.8	52.3	Kurgan and Homaeian [31]
25PDB	StackingC ensemble	34 features	87.6	59.9	This paper
	Logistic regression	66 features	62.2	57.1	Kurgan and Homaeian [31]

Comparison with competing methods

The three standard benchmarking datasets, i.e., “359”, “1189”, and 25PDB, were used to compare the proposed method with other competing structural class prediction methods. The jackknife and resubstitution results together with the corresponding references are shown in Table 5.

For the smallest, highly homologous “359” dataset our methods achieved one of the highest results, which is only 0.6% lower than results of Kurgan and Homaeian [31] who used twice as many features to represent the protein sequences. At the same time, our accuracy for this dataset is higher than accuracy of all other methods that used a similar number of features. For the “1189” dataset, which includes sequences of up to 40% homology, the proposed method obtained 5% improvement in accuracy (10.8% reduction of error rate) when compared with the two other papers. Finally, results on the 25PDB dataset again show superiority of the proposed method; 2.8% improvement (6.5% reduction of error rate) was achieved when compared to results in [31] despite using half as many features. In short, the results show that the proposed structural class prediction method improves accuracy for almost all considered datasets, while applying a compact feature representation.

Summary and conclusion

This paper proposes a novel method for prediction of structural classes based on protein sequences. The method applies a custom designed feature-based representation of the sequences and an ensemble of four complementary classifiers to improve prediction accuracy for sequences of varying homology. Two classifiers are characterized by relatively high accuracy for low-homology sequences, while the other two by high accuracy for high homology sequences. Four controlled homology (25%, 50%, 70%, and 90%) datasets, which were extracted from the PDB, were used to select the best ensemble method and a compact feature representation. Experimental tests have shown that the application of the new sequence representation (which includes 34 features that describe AA composition, polarity, electric charge, and hydrophobicity) and the stacking ensemble gave on average 9.8% increase in accuracy (on average the error rates were reduced by 22.3%) when compared with the commonly used composition vector representation and the best individual classifiers. The error rates were reduced by 21.2%, 21.4%, 22.3%, and 24.4% in function of increasing homology for the four controlled homology datasets, showing that the improvements are relatively large across the entire spectrum of sequence similarity levels.

The proposed prediction method was compared with other competing prediction methods, which include support vector machines, regression, and geometric classifiers, on three benchmark datasets. Our method was superior to most other methods, with the exception of result for the most homologous dataset given in [31]. The prediction accuracy ranges between 60% for the low-homology sequences and over 73% accuracy for the high homology sequences. To compare, the latest competing methods report 52% accuracy [44] and 57% accuracy [31] for analogous problems.

The results suggest that application of an ensemble of classifiers together with appropriate feature representation has a high potential to provide accurate prediction of protein structural classes.

Acknowledgment

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

References

[1] D. Aha, D. Kibler, Instance-based learning algorithms, Machine Learning 6 (1991) 37–66.

[2] A. Andreeva, D. Howorth, S. Brenner, T. Hubbard, C. Chothia, A. Murzin, SCOP database in 2004: refinements integrate structure and sequence family data, Nucleic Acid Research 32 (2004) D226–D229.

[3] I. Bahar, A.R. Atilgan, R.L. Jernigan, B. Erman, Understanding the recognition of protein structural classes by amino acid composition, Proteins 29 (1997) 172–185.

[4] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I. Shindyalov, P. Bourne, The Protein Data Bank, Nucleic Acids Research 28 (2000) 235–242.

[5] S. Black S, D. Mould, Development of hydrophobicity parameters to analyze proteins which bear post- or cotransitional modifications, Analytical Biochemistry 193 (1991) 72–82.

[6] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

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

[8] Y. Cai, Is it a paradox or misinterpretation? Proteins 43 (2001) 336–338.

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

[10] O. Camoglu, T. Can, A.K. Singh, Y.F. Wang, Decision tree based information integration for automated protein classification, Journal of Bioinformatics and Computational Biology 3 (3) (2005) 717–742.

[11] K.C. Chou, C.T. Zhang, A new approach to predicting protein folding types, Journal of Protein Chemistry 12 (1993) 169–178.

[12] K.C. Chou, C.T. Zhang, Predicting protein-folding types by distance functions that make allowances for amino-acid interactions, Journal of Biological Chemistry 269 (1994) 22014–22020.

[13] K.C. Chou, A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space, Proteins 21 (1995) 319–344.

[14] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, Critical Review in Biochemistry and Molecular Biology 30 (1995) 275–349.

[15] K.C. Chou, G.M. Maggiora, Domain structural class prediction, Protein Engineering 11 (1998) 523–538.

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

[17] J. Cornette, K.B. Cease, H. Margalit, J.L. Spouge, J.A. Berzofsky, C. DeLisi, Hydrophobicity scales and computational techniques for detecting amphipathic structures in protein, Journal of Molecular Biology 195 (1987) 659–685.

[18] S. Diplaris, G. Tsoumakas, P. Mitkas, I. Vlahavas, Protein Classification with Multiple Algorithms, Proceedings of the 10th Panhellenic Conference on Informatics, 2005, 448–456.

[19] I. Dubchak, I. Muchnik, S.R. Holbrook, S.H. Kim, Prediction of protein-folding class using global description of amino-acid sequence, Proceedings of the National Academy of Science 92 (1995) 8700–8704.

[20] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, S.H. Kim, Recognition of a protein fold in the context of the SCOP classification, Proteins 35 (1999) 401–407.

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

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

[23] M. Gromiha, S. Selvaraj, Protein secondary structure prediction in different structural classes, Protein Engineering 11 (1998) 249–251.

[24] M.A. Hall, Correlation-Based Feature Subset Selection for Machine Learning, Ph.D. thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1999.

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

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

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

[28] K. Kedarisetti, L. Kurgan, S. Dick, A comment on "Prediction of protein structural classes by a new measure of information discrepancy", Computational Biology and Chemistry (2006), accepted.

[29] P. Klein, C. Delisi, Prediction of protein structural class from the amino-acid sequence, Biopolymers 25 (1986) 1659–1672.

[30] R. Kohavi, G. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1-2) (1997) 273–324.

[31] L. Kurgan, L. Homaeian, Prediction of Structural Classes for Protein Sequences and Domains—Impact of Prediction Algorithms, Sequence Representation and Homology, and Test Procedures on Accuracy, Pattern Recognition, special issue on Bioinformatics, published online on April 11 2006.

[32] W. Li, L. Jaroszewski, A. Godzik, Tolerating some redundancy significantly speeds up clustering of large protein databases, Bioinformatics 18 (1) (2002) 77–82.

[33] Z. Lin, X-M. Pan, Accurate prediction of protein secondary structural content, Journal of Protein Chemistry 20 (3) (2001) 217–220.

[34] H. Liu and R. Setiono, A Probabilistic Approach to Feature Selection—A Filter Solution, Proceedings of the 13th International Conference on Machine Learning, 1996, 319–327.

[35] S. le Cessie, van J.C. Houwelingen, Ridge estimators in logistic regression, Applied Statistics 41 (1) (1992) 191–201.

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

[37] R. Luo, Z. Feng, J. Liu, Prediction of protein structural class by amino acid and polypeptide composition, European Journal of Biochemistry 269 (2002) 4219–4225.

[38] B. Metfessel, P. Saurugger, P. Connelly, S. Rich, Cross-validation of protein structural class prediction using statistical clustering and neural networks, Protein Science 2 (1993) 1171–1182.

[39] A. Murzin, S. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of protein database for the investigation of sequence and structures, Journal of Molecular Biology 247 (1995) 536–540.

[40] H. Nakashima, K. Nishikawa, T. Ooi, The folding type of a protein is relevant to the amino acid composition, Journal of Biochemistry 99 (1986) 153–162.

[41] J. Platt, Fast training of support vector machines using sequential minimal optimization, in: B. Schoelkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods—Support Vector Learning, MIT Press, Cambridge, MA, 1998.

[42] A.K. Seewald, How to make stacking better and faster while also taking care of an unknown weakness, in: C. Sammut, A. Hoffmann (Eds.), Proceedings of the 19th International Conference on Machine Learning (ICML 2002), 2002, pp.554–561.

[43] H.B. Shen, J. Yang, X-J. Liu, K.C. Chou, Using supervised fuzzy clustering to predict protein structural classes, Biochemical and Biophysical Research Communications 334 (2005) 577–581.

[44] X-D. Sun, R.B. Huang, Prediction of protein structural classes using support vector machines, Amino Acids 30 (2006) 469–475.

[45] A.C. Tan, D. Gilbert, Y. Deville, Multi-class protein fold classification using a new ensemble machine learning approach, Genome Informatics 14 (2003) 206–217.

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

[47] I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, second ed., Morgan Kaufmann, San Francisco, 2005.

[48] C.T. Zhang, K.C. Chou, An optimization approach to predicting protein structural class from amino-acid composition, Protein Science 1 (1992) 401–408.

[49] C.T. Zhang, K.C. Chou, G.M. Maggiora, Predicting protein structural classes from amino acid composition: application of fuzzy clustering, Protein Engineering 8 (1995) 425–435.

[50] Z.D. Zhang, Z.R. Sun, C.T. Zhang, A new approach to predict the Helix/strand content of globular proteins, Journal of Theoretical Biology 208 (2001) 65–78.

Please cite this article as: Kanaka Durga Kedarisetti et al., Classifier ensembles for protein structural class prediction ..., Biochemical and Biophysical Research Communications (2006), doi:10.1016/j.bbrc.2006.07.141.