BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 21 2007, pages 2843–2850
doi:10.1093/bioinformatics/btm475

Structural bioinformatics

PFRES: protein fold classification by using evolutionary information and predicted secondary structure

Ke Chen and Lukasz Kurgan*

Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada

Received on May 28, 2007; revised on August 27, 2007; accepted on September 17, 2007

Advance Access publication October 17, 2007

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: The number of protein families has been estimated to be as small as 1000. Recent study shows that the growth in discovery of novel structures that are deposited into PDB and the related rate of increase of SCOP categories are slowing down. This indicates that the protein structure space will be soon covered and thus we may be able to derive most of remaining structures by using the known folding patterns. Present tertiary structure prediction methods behave well when a homologous structure is predicted, but give poorer results when no homologous templates are available. At the same time, some proteins that share twilight-zone sequence identity can form similar folds. Therefore, determination of structural similarity without sequence similarity would be beneficial for prediction of tertiary structures.

Results: The proposed PFRES method for automated protein fold classification from low identity (<35%) sequences obtains 66.4% and 68.4% accuracy for two test sets, respectively. PFRES obtains 6.3–12.4% higher accuracy than the existing methods. The prediction accuracy of PFRES is shown to be statistically significantly better than the accuracy of competing methods. Our method adopts a carefully designed, ensemble-based classifier, and a novel, compact and custom-designed feature representation that includes nearly 90% less features than the representation of the most accurate competing method (36 versus 283). The proposed representation combines evolutionary information by using the PSI-BLAST profile-based composition vector and information extracted from the secondary structure predicted with PSI-PRED.

Availability: The method is freely available from the authors upon request.

Contact: lkurgan@ece.ualberta.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Protein structures are being solved to answer key biological questions related to protein function, regulation and interactions. Outside of their biological context, the solved structures are increasingly useful for structure modeling/prediction for unsolved protein sequences that have a closely related (similar) sequence with known structure (Tress *et al*., 2005; Wang *et al*., 2005). Based on the Chothia's estimate, which states that the number of different protein families is finite and perhaps as small as 1000 (Chothia, 1992), it seems feasible to derive most of the unsolved structures by homology modeling based only on a relatively small portion of the protein structures that are determined experimentally. This explains why the novel structures are especially valuable. This fact also served as the basis of the Protein Structure Initiative that was initiated by NIH in 1999 (Chandonia and Brenner, 2006). One of the aims of this project is to cover the structure space of proteins. These early findings are supported by a recent computational analysis of the Protein Data Bank, which showed that the growth of the structural data has slowed down and that the rate of increase of the related SCOP categories (including number of families, superfamilies and folds) is also slowing down (Levitt, 2007). Homology modeling is based on the assumption that homologous sequences share similar folding patterns (Ruan *et al*., 2006; Zhang and Skolnick, 2005). At the same time, sequences with low sequence identity can also share similar folding patterns (Paiardini *et al*., 2004) and can be used to predict tertiary structure (Bujnicki, 2006). Sequence alignment software is an important tool to find homologous sequences among the known structures (Altschul *et al*., 1997; Yu *et al*., 2006), but inept when no homologous sequences are available. Research also shows that finding similar folding patterns among the low identity sequences is beneficial for reconstruction of the tertiary structure (Reinhardt and Eisenberg, 2004; Tomii *et al*., 2005).

A comprehensive and detailed description of the structural relationships between all solved proteins is provided in the SCOP (Structural Class of Proteins) database (Andreeva *et al*., 2004; Murzin *et al*., 1995). This database implements a hierarchy of relations between known protein and protein domain structures. The classification on the first level of the hierarchy is commonly known as the protein structural class, while the second level classifies proteins into folds, which are the classification target in this article. Several machine-learning methods have been applied to detect the structurally similar proteins (protein folds) from sequences that share low identity. Ding and Dubchak investigated support vector machine (SVM) and neural network for protein fold classification (Ding and Dubchak, 2001). Shen and Chou studied ensemble models based on nearest neighbor (Shen and Chou, 2006). The prediction of the type of the protein fold for a given sequence

*To whom correspondence should be addressed.

is performed with an intermediate step that converts the sequence into a feature space representation. Several other ensemble models that applied the same feature space representation as the one proposed by Ding and Dubchak were also proposed (Bologna and Appel, 2002; Nanni, 2006; Okun, 2004). In these studies protein sequences were represented by composition vector (CV), predicted secondary structure, hydrophobicity, normalized van der Waals volume, polarity, polarizability and pseudo-amino acid composition. The fold classification success rate ranged between 56% and 62%. The dimensionality of the feature space was relatively high, i.e. 125 features were proposed by Ding and Dubchak and 283 features by Chou and Shen, when compared with size of the dataset used in the experimental evaluation, i.e. 313 training and 385 test proteins. To this end, we propose a novel fold classification method, called PFRES, that provides significantly better prediction accuracy when compared with the existing methods and that uses a small number of new and more effective features. The main source of the achieved improvement is attributed to the application of PSI-BLAST profile (Altschul *et al.*, 1997) based composition vector, which considers evolutionary information (Jones, 1999, 2007; Kim and Park, 2004), instead of the composition and pseudo-composition vectors that were used in the prior works. We also applied features generated from secondary structure predicted with PSI-PRED (Jones, 1999), which are also shown to be beneficial in the context of the fold classification. Finally, we note that PFRES, as well as all other relevant competing methods, address a simplified fold classification problem, i.e. they predict 27 folds, due to low counts of proteins that belong to the remaining folds.

2 MATERIALS AND METHODS

2.1 Datasets

The proposed method was designed on a training dataset with 313 domains proposed by Ding and Dubchak (Ding and Dubchak, 2001). The tests were performed on two datasets: the test set 1 with 385 domains was also taken from Ding and Dubchak (2001) and was used to perform comparison with other existing methods; the test set 2 with 908 domains was included to provide larger scale evaluation on more recently deposited domains and to assure that the proposed method does not overfit the first, small test set.

The follow-up study by Shen and Chou excluded two training domains (2SCMC and 2GPS) and two domains from test set 1 (2YHX_1 and 2YHX_2) due to lack of sequence records (Shen and Chou, 2006). We follow Shen and Chou's study and adopt the two datasets without these four sequences. The sequence identity for any pair of sequences in the training set is <35%. According to the dataset authors, the sequence in test set 1 share more than 35% sequence identity with the sequences in the training set. We found seven duplicates between these two sets, i.e. 1APLC, 3RUB2, 2REB1, 1DSBA2, 1GLCG1, 1GLCG2 and 1SLTA from the training set correspond to 1YRNB, 3RUBL2, 2REB_1, 1DSBA2, 1GLAG1, 1GLAG2 and1SLAA from the test set 1, respectively. We also found another 12 pairs that share over 50% identity. This redundancy may result in overestimated test results on dataset 1, but at the same time it should not impact ability to compare the relative differences between prediction accuracies achieved by various methods on this test set. The training and test set 1 sequences belong to the following 27 folds: (1) globin-like, (3) cytochrome c, (4) DNAbinding3-helical bundle, (7) 4-helical up-and-down bundle, (9) 4-helicalcytokines, (11) EF-hand, (20) immunoglobulin-like, (23) cupredoxins, (26) viral coat and capsid proteins, (30) conA-like lectin/glucanases, (31) SH3-like barrel, (32) OB-fold, (33) beta-trefoil, (35) trypsin-like serine proteases, (39) lipocalins, (46) (TIM)-barrel, (47) FAD (also NAD)-binding motif, (48) flavodoxin-like, (51) NAD(P)-binding Rossmann-fold, (54) P-loop, (57) thioredoxin-like, (59) ribonuclease H-like motif, (62) hydrolases, (69) periplasmic binding protein-like, (72) b-grasp, (87) ferredoxin-like and (110) small inhibitors, toxins and lectins. These 27 folds are the most populated in SCOP; each of them contains at least seven proteins. Based on the concept of protein structural class proposed by Levitt and Chothia (Levitt and Chothia, 1976), folds 1–11 belong to all-α structural class, folds 20–39 to all-β class, folds 46–69 to α/β class and folds 72–87 to $\alpha+\beta$ class. The fold distribution can be found in Ding and Dubchak (2001) and these two datasets can be downloaded from Supplementary Material in Shen and Chou (2006).

Test set 2 includes sequences that belong to the same 27 folds and that were deposited into PDB between 2002 and 2004. The selected timeframe is a result of two factors: the newest version of SCOP assigned folds only for sequences deposited until January 2005, while the training set and test set 1 were generated before 2001 and we aimed to avoid overlap between these datasets. The sequences in test set 2 were filtered by CD-HIT (Li and Godzik, 2006) at 40% sequence identity. Next, the remaining sequences were aligned with the sequences in both the training set and test set 1 using Smith–Waterman algorithm (Smith and Waterman, 1981). Only sequences that have <35% sequence identity with any sequence in these two sets were selected to form the test set 2. The resulting 908 sequences are available from the authors upon request.

2.2 Feature space representation

2.2.1 PSI-BLAST profile-based composition vector The composition vector (*CV*) is computed directly from amino acid (AA) sequence (Chen *et al.*, 2007; Chou, 2005). Given that the 20 AAs, which are ordered alphabetically (A, C, ..., W, Y), are represented as AA_1, AA_2, ..., AA_{19} and AA_{20}, and the number of occurrences of AA_i in the entire sequence is denoted as n_i, the composition vector is defined as:

$$\left(\frac{n_1}{L}, \frac{n_2}{L}, \ldots, \frac{n_{19}}{L}, \frac{n_{20}}{L}\right)$$

where L is the length of the sequence. This representation was used by majority of the existing fold classification methods (Bologna and Appel, 2002; Ding and Dubchak, 2001; Nanni, 2006; Okun, 2004).

The new representation, which combines PSI-BLAST profile and the concept of composition vector, was developed for the proposed prediction method. The prior successful applications of PSI-BLAST profile illustrate that the evolutionary information is more informative than the query sequence itself (Jones, 1999, 2007; Kim and Park, 2004). PSI-BLAST aligns a given query sequence to a database of sequences. Using multiple sequence alignment, PSI-BLAST counts the frequency of each AA at each position for the query sequence and generates 20-dimensional vector of AA frequencies for each position in the query sequence. The generated PSI-BLAST profile can be used to identify key positions of conserved AAs and positions that undergo mutations. Our approach combines the composition vector of the entire sequence and the PSI-BLAST profile into so called *PSI-BLAST profile-based composition vector* (*PCV*). The PSI-BLAST profile is an $L \times 20$ matrix, which is denoted as $[a_{i,j}]$, where $i = 1,2,\ldots,L$ denotes position in the query sequence and $j = 1,2,\ldots,20$ denotes a given AA. After applying the substitution matrix and log function, a_{ij} values range between -9 and 11. The proposed representation is related to calculation of the composition vector based on binary coding. The binary coding uses a 20-dimensional vector to encode each AA. In binary coding, AA_i is encoded as $(0,0,\ldots,0,1,0,\ldots,0,0)$, where only the ith value is greater

than 0. The binary coding matrix is denoted as $[b_{i,j}]$. The binary encoding and PSI-BLAST profile matrices have the same dimensionality ($L \times 20$).

CV can be computed from the binary coding matrix in a straightforward way. For a given protein sequence $A_1A_2 \ldots A_N$

$$CV_i = \sum_{k=1}^{L} \frac{b_{ki}}{L} \quad (i = 1,2,\ldots,20)$$

where $\{CV_i,\ i=1,2,\ldots,20\}$ is the 20-dimensional composition vector.

PCV is calculated in a similar way. The only difference is that the binary coding matrix $[b_{i,j}]$ is replaced by PSI-BLAST profile $[a_{i,j}]$. Therefore, *PCV* is defined as:

$$PCV_i = \sum_{k=1}^{L} \frac{a_{ki}}{L} \quad (i = 1,2,\ldots,20)$$

Since PSI-BLAST profile values can be negative, while the frequencies of AA pairs should not be negative, we redefine *PCV* as follows:

$$PCV_i = \sum_{k=1}^{L} \frac{\max(a_{ki}, 0)}{L} \quad (i = 1,2,\ldots,20)$$

where the negative a_{ki} values are replaced by 0 and the 20-dimensional $\{PCV_i,\ i=1,2,\ldots,20\}$ vector corresponds to the PSI-BLAST profile-based composition vector.

2.2.2 Secondary structure predicted with PSI-PRED Predicted secondary structure is proven to be helpful in fold classification. The recently proposed fold classification studies (Ding and Dubchak, 2001; Shen and Chou, 2006) used the secondary structure predicted with relatively older methods (Holley and Karplus, 1989; Quian and Sejnowski, 1988). In contrast, we use a more recent PSI-PRED method (Jones, 1999), which is shown to provide superior accuracy when compared with other state-of-the-art competing secondary structure prediction methods (Birzele and Kramer, 2006; Lin *et al.*, 2005). We used PSI-PRED25 with default parameters to predict secondary structure from the protein sequences. The 3-state predictions (helix, strand and coil) are used to generate the features.

Secondary structure content (*SSC*) is shown to improve classification accuracy of a related problem of structural class prediction (Kurgan and Chen, 2007). To this end, we introduce *SSC* that is calculated from the secondary structure predicted with PSI-PRED. Let us denote the AA sequence as $\{A_i,\ i=1,2,\ldots,L\}$ and the predicted secondary structure as $\{S_i,\ i=1,2,\ldots,L\}$. We count the occurrences of 'H', 'E' and 'C' predictions and denote the corresponding counts as $\mathrm{COUNT_H}$, $\mathrm{COUNT_E}$, $\mathrm{COUNT_C}$, respectively. The *SSC* is defined as:

$$Content_{class} = \frac{COUNT_{class}}{L}$$

where *class* = 'H', 'E' and 'C'.

Number of distinct secondary structure segments (*DSSS*). Although secondary structure content reflects information about the secondary structure of the entire sequence, it does not provide information concerning individual secondary structure segments. At the same time, size (length) of secondary structure segments is one of the deciding factors when it comes to the classification of the structural classes and folds. To this end, we designed features that count the number of occurrences of distinct helix, strand and coil structures which length (number of the corresponding AAs) is above a certain threshold. In this way short secondary structure segments, which possibly can be incorrectly predicted, will be filtered out. We varied the threshold values between 2 and 9 for the strand and coil segments and between 3 and 9 for the helix segments and run predictions using SVM classifier. The corresponding results are shown in Figure 1. Based on the graph, the threshold to count helical segments equals 7. The thresholds for strand and coil segments equal 5 and 6, respectively. We note that the accuracies resulting from using different threshold are relatively similar, i.e. within 1%, and thus the quality of the proposed method should not be sensitive to this parameter.

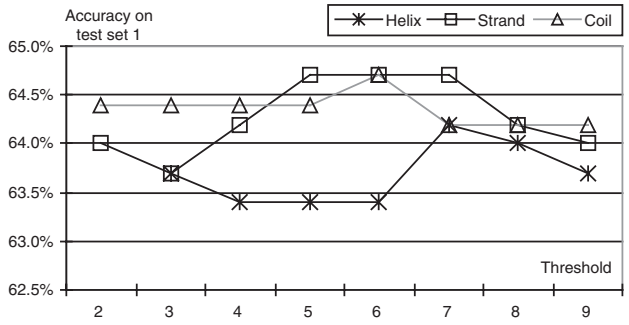

Fig. 1. Optimization of segment length thresholds to define DSSS features.

Table 1. Summary of the feature selection results

Features set	Total number features	Selected features
PCV	20	20
SSC	3	3
Number of *DSSS*	3	2
Arrangement of *DSSS*	27	10
Length	1	1
Total	54	36

Arrangement of DSSS. In some cases, structural folds cannot be distinguished based on the *SSC* and *DSSS* features. For instance, the α/β and $\alpha+\beta$ classes contain both α-helices and β-strands; the α/β class includes mainly parallel β-strands, while $\alpha+\beta$ class mainly includes anti-parallel strands, which is related to the arrangement of secondary structure segments, but not the *SSC* or *DSSS* values. Therefore, we also designed another set of features that encode arrangement of three neighboring secondary structure segments, which meet the minimum threshold criteria set for *DSSS* features. There are 27 possible segment arrangements, i.e. *class-class-class* where *class* = 'H', 'E' and 'C'. We count the corresponding number of occurrences for each arrangement.

Finally, we also include the length of the sequence (L) as a feature. Table 1 summarizes features used in this article.

2.3 Feature selection

Feature selection method was used to reduce the dimensionality and potentially improve the prediction accuracy. An entropy-based feature selection method (Yu and Liu, 2003), which evaluates each feature by measuring the information gain with respect to the class (protein fold), was applied.

The entropy of a feature X is defined as:

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

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

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

where $P(x_i \mid y_j)$ is the posterior probability of X given the value y_i of Y.

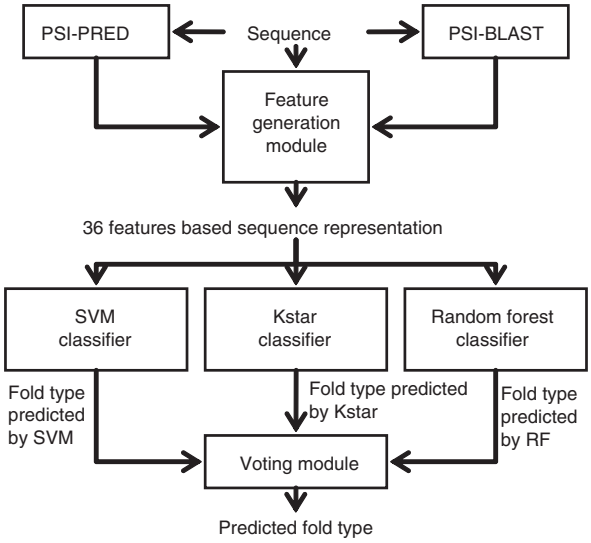

Fig. 2. Architecture of the proposed fold classification method.

The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called information gain

$$IG(X|Y) = H(X) - H(X|Y)$$

According to this measure, Y has stronger correlation with X than with Z if $IG(X|Y) > IG(Z|Y)$. The feature selection was performed using 10-fold cross-validation on the training set. Among the original set of 54 features, 36 with the best information gain values were selected; see Table 1.

2.4 Proposed prediction method

The proposed prediction method was designed and tested in two steps. First, we selected a set of best-performing classifiers among six state-of-the-art methods that include SVM (Kerthi *et al.*, 2001), Multiple Logistic Regression (Le and Houwelingen, 1992), instance learning-based Kstar (Cleary and Trigg, 1995) and IB1 (Aha and Kibler, 1991) algorithms, Naïve Bayes (John and Langley, 1995), and Random Forest (Leo, 2001) and when using the selected 36 features to represent sequences. Second, three different ensembles of the selected classifiers, including voting, grading and stacking (Seewald, 2002; Seewald and Fuernkranz, 2001), were tried and the best performing ensemble was used to implement our fold classification method. As a result, voting-based ensemble, which combines predictions from the three classifiers based on an unweighted average of the corresponding classification probability estimates, was selected. The architecture of the proposed PFRES method is shown in Figure 2. The classification algorithms used to develop and compare the proposed method were implemented in Weka (Witten and Frank, 2005).

3 RESULTS AND DISCUSSION

The experiments first report results related to the design of the proposed fold classification method. We also test and discuss effectiveness of individual feature sets from the proposed sequence representation. Finally, the results of our ensemble method are compared with the results of five competing methods.

For the test set 1, the fold classification accuracies for the six classifiers that include SVM, Multiple Logistic Regression, Kstar, IB1, Naïve Bayes and Random Forest and when using the selected 36 features to represent sequences are shown in Table 2. Random Forest (with 250 trees) gives the highest accuracy, i.e. 66.8%, among the six classifiers. The two runner-up classifiers, SVM (with RBF kernel with $\gamma = 0.8$ and complexity parameter $C = 5.0$) and Kstar (with global blend = 96), obtained 66.1% and 65.0% accuracy, respectively. The same classifiers were also evaluated on the test set 2 by applying the same group of features and the same parameters. Random Forest again gives the highest accuracy, i.e. 63.3%, with the same two runner-up classifiers, SVM and Kstar, which obtained 62.4% and 62.7% accuracy, respectively. The accuracies on test set 2 are slightly lower than accuracies on test set 1. The remaining three classifiers obtained accuracy that is 3–10% lower than the accuracy of the three best classifiers, and thus were not used to implement the proposed fold classification method.

Among the 27 folds, fold 1 and fold 39 are the easiest to classify, i.e. all six classifiers achieved 100% accuracy for these two folds. Folds 3, 7, 9, 26, 33, 47 and 110 are also relatively easy to classify, i.e. the average accuracy of the six classifiers for these folds is above 80%. The average prediction accuracy for all-α structural class (folds 1–11) is 77.1%, for all-β class (folds 20–39) is 64.3%, for α/β class (folds 46–69) is 55.6% and for $\alpha + \beta$ class (folds 72–87) is 40%. The folds that belong to all-α and all-β structural classes are easier to classify, while folds that belong to α/β and $\alpha + \beta$ classes are more difficult to correctly recognize. This is expected as the proposed features, and especially those based on the predicted secondary structure, should be able to successfully represent proteins that contain mainly α-helices and β-strands. At the same time, although still well performing, the proposed features are less efficient in capturing long range interactions that are characteristic to formation of parallel and anti-parallel β-strands.

3.1 Effectiveness of PCV features

The PSI-BLAST profile-based composition vector (*PCV*), which is proposed in this article, was directly compared with the corresponding sequence-based composition vector (*CV*) representation that was used in Bologna and Appel (2002), Ding and Dubchak (2001), Nanni (2006) and Okun (2004). *PCV* and *CV* were compared based on fold classification performed with the six classifiers. The prediction results are shown in Figure 3. The comparison shows consistent superior quality of PCV features, i.e. the results based on *PCV* features are at least 13% higher than the result from *CV* features for all six classifiers. For test set 1, the average accuracy when using *PCV* features is 54.8%, while for *CV* features it drops to 39.1%. For the test set 2, the average accuracy when using *PCV* features is 46.3%, while for *CV* features it drops to 27.3%. This illustrates that sequential evolutionary information is critical for successful classification of protein folds, even for sequences that share low sequence identity. The results also indicate that the test set 2 is more challenging.

Table 2. Comparison of prediction accuracies between different classifiers for the proposed sequence representation that includes the selected 36 features

Folds	Individual classifiers						Ensemble classifiers		
	SVM	Kstar	Random Forest	IB1	Naïve Bayes	Regression	Grading	Voting	Stacking-C
1	100	100	100	100	100	100	100	100	100
3	100	100	100	88.9	100	100	100	100	100
4	45	45	70	40	65	50	55	60	60
7	100	62.5	100	87.5	100	75	100	75	100
9	100	88.9	88.9	88.9	100	100	88.9	88.9	88.9
11	77.8	66.7	66.7	66.7	66.7	66.7	66.7	66.7	66.7
20	75	84.1	77.3	65.9	52.3	65.9	86.4	81.8	79.5
23	33.3	16.7	33.3	25	33.3	41.7	25	33.3	33.3
26	84.6	100	92.3	84.6	92.3	84.6	100	92.3	92.3
30	66.7	66.7	83.3	66.7	50	66.7	66.7	66.7	83.3
31	50	62.5	37.5	37.5	50	62.5	50	62.5	37.5
32	52.6	47.4	52.6	47.4	63.2	21.1	52.6	52.6	63.2
33	100	75	50	100	100	100	75	75	75
35	25	50	50	50	25	25	50	50	50
39	100	100	100	100	100	100	100	100	100
46	64.6	66.7	66.7	62.5	39.6	47.9	68.8	68.8	66.7
47	83.3	91.7	83.3	83.3	83.3	75	83.3	91.7	91.7
48	30.8	30.8	46.2	38.5	53.8	30.8	38.5	46.2	38.5
51	59.3	70.4	59.3	70.4	48.1	29.6	63	66.7	66.7
54	50	41.7	33.3	41.7	41.7	41.7	41.7	33.3	33.3
57	37.5	50	37.5	37.5	37.5	62.5	50	50	37.5
59	66.7	58.3	66.7	66.7	58.3	58.3	58.3	66.7	66.7
62	57.1	57.1	42.9	85.7	42.9	57.1	42.9	57.1	57.1
69	50	25	50	50	50	25	50	50	50
72	25	25	25	25	25	25	25	25	25
87	55.6	40.7	51.9	33.3	51.9	33.3	48.1	51.9	51.9
110	96.3	88.9	96.3	88.9	92.6	66.7	96.3	96.3	96.3
Overall	66.1	65	66.8	62.1	60.3	55.1	67.6	68.4	68.1

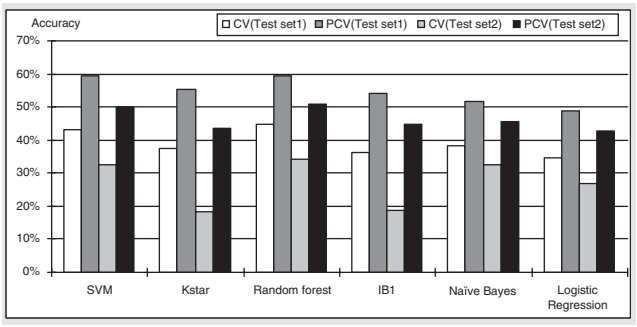

Fig. 3. Comparison of prediction accuracies (y axis) between PSI-BLAST profile-based composition vector and sequence-based composition vector. Two sets of feature were tested with six classifiers (x axis) on both test sets.

3.2 Effectiveness of features based on the predicted secondary structure

Features generated from the predicted secondary structure that were proposed in this article, which include *SSC*, number of *DSSS* and the arrangement of *DSSS*, are also shown to contribute to the improved fold classification. We compared prediction accuracy of the three best classifiers when using the *PCV* features with accuracy when the features computed from the predicted secondary structure are added, see Figure 4. For test set 1, 55.4%, 59.5% and 59.3% accuracies were obtained for Kstar, Random Forest and SVM classifiers, respectively, when using only *PCV* to represent sequences. After adding *SSC* features, the accuracies increased to 57.7%, 62.4% and 60.1%. By adding the number of *DSSS*, the accuracies again increased to 61.4%, 65.8% and 64.8%. Finally, adding the features related to the arrangement of *DSSS* results in accuracies of 63.4%, 65.8% and 65.5%. Similar results were observed for the test set 2. The accuracies of Kstar, Random Forest and SVM classifiers equal 43.6%, 50.9% and 50% when using only *PCV*, 49.7%, 57.3% and 57.9% after adding *SSC* features, 55.7%, 63% and 61.5% after adding the number of *DSSS*, and finally 61%, 63% and 62.6% after adding the features related to the arrangement of *DSSS*, respectively. These consistent improvements show that each of the proposed features sets results in improvements and

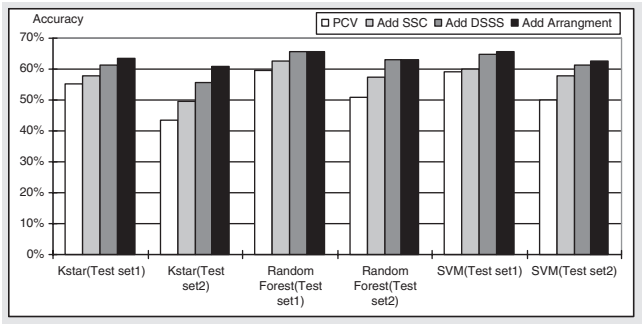

Fig. 4. Comparison of classification accuracy (*y* axis) obtained by using features calculated from the secondary structure predicted by PSI-PRED, i.e. *PCV* features only, *PCV* and *SSC* features, *PCV*, *SSC* and number of *DSSS* features, and *PCV*, *SSC*, number of *DSSS* and arrangement of *DSSS* features. Results of the three best classifiers on both test sets (*x*axis) are shown.

illustrates the importance of the secondary structure information with respect to the classification of protein folds.

3.3 Comparison of ensemble models

Several prior works on protein folds classification applied ensemble models to improve prediction accuracy (Bologna and Appel, 2002; Nanni, 2006; Shen and Chou, 2006). The method by Shen and Chou (Shen and Chou, 2006) ensembles nine evidence-theoretic *k*-nearest neighbor classifiers that use different input feature sets. The ensemble proposed in Bologna and Appel (2002) applies four specialized neural networks that use different subsets of protein sequences from the training set. Finally, the ensemble developed by Nanni (Nanni, 2006) uses 27 *k*-local hyperplane-based nearest neighbor classifiers, each of which uses different subset of features among these proposed in Ding and Dubchak (2001). In contrast to the above methods that ensemble the same type of classifiers, our method ensembles three different classifiers that provide complementary predictions, i.e. SVM provides superior predictions for folds 9, 11, 33, 54 and 87; Kstar for folds 20, 26, 31, 47, 51 and 57; Random Forest for folds 4, 30 and 48; see Table 2. Three methods for combining multiple classifiers that include voting, grading and stacking were compared on test set 1; see Table 2. All three ensembles are shown to provide better accuracies than the best single classifier, Random Forest. The proposed method adopts the best performing voting-based ensemble that achieves 68.4% accuracy on test set 1. For the test set 2, the same voting-based ensemble achieves 66.4% accuracy. In case of both test sets, folds 1, 3, 9, 20 and 110 were predicted with accuracy of above 80%, while accuracy of below 50% was recorded for folds 23, 35, 48 and 72. Results on both test sets show that the application of the ensemble model results in 2–3% improvement in prediction accuracy over the prediction based on single classifier. The lower prediction accuracy on the test set 2 could be explained by the strict separation (up to 35% sequence similarity) between this test set and the training set. In contrast, test set 1 is shown to share some redundant and similar sequences with the training set. When these 19 sequences were removed from test set 1, the PFRES obtains 67% accuracy on this set, which is only 0.6% higher than accuracy on the test set 2.

Table 3. Comparison between PFRES and the competing fold classification methods on test set 1. The best results for each fold are shown in bold

Folds	Fold classification methods					
	SVM[a] (%)	HKNN[b] (%)	DIMLP[c] (%)	SE[d] (%)	PFP[e] (%)	PFRES this article
1	83.3	83.3	85.0	83.3	83.3	**100**
3	77.8	77.8	97.8	88.9	55.6	**100**
4	35.0	50.0	66.0	70.0	**85.0**	60.0
7	50.0	**87.5**	41.3	50.0	75.0	75.0
9	**100**	88.9	91.1	**100**	**100**	88.9
11	**66.7**	44.4	22.2	33.3	33.3	**66.7**
20	71.6	56.8	75.7	79.6	70.5	**81.8**
23	16.7	25.0	**40.0**	25.0	16.7	33.3
26	50.0	84.6	80.8	69.2	**100**	92.3
30	33.3	50.0	46.7	33.3	33.3	**66.7**
31	50.0	50.0	**75.0**	62.5	37.5	62.5
32	26.3	42.1	22.6	36.8	15.8	**52.6**
33	50.0	50.0	45.0	50.0	**75.0**	**75.0**
35	25.0	**50.0**	**50.0**	25.0	**50.0**	**50.0**
39	57.1	42.9	74.3	28.6	71.4	**100**
46	77.1	79.2	83.8	87.5	**97.9**	68.8
47	58.3	58.3	55.0	58.3	66.7	**91.7**
48	48.7	53.9	52.3	**61.5**	15.4	46.2
51	61.1	40.7	39.3	37.0	44.4	**66.7**
54	36.1	33.3	41.7	**50.0**	33.3	33.3
57	50.0	37.5	46.3	50.0	**62.5**	50.0
59	35.7	**71.4**	55.0	64.3	66.7	66.7
62	**71.4**	**71.4**	44.3	**71.4**	57.1	57.1
69	25.0	25.0	25.0	25.0	**50.0**	**50.0**
72	12.5	25.0	23.8	25.0	**37.5**	25.0
87	37.0	25.9	41.1	33.3	29.6	**51.9**
110	83.3	85.2	**100**	85.2	96.3	96.3
Overall	56.0	57.1	61.1	61.1	62.1	**68.4**

[a]refers to (Ding and Dubchak, 2001).
[b]refers to (Okun, 2004).
[c]refers to (Bologna and Appel, 2002).
[d]refers to (Nanni, 2006).
[e]refers to (Shen and Chou, 2006).

3.4 Comparison with competing prediction methods

The proposed PFRES method was compared with five recent methods that address same task on test set 1; see Table 3. Ding and Dubchak's method uses representation with 125 features and SVM and neural networks as the classifiers (Ding and Dubchak, 2001). Okun's method uses features proposed in Ding and Dubchak (2001) and *k*-local hyperplane nearest neighbor classifier (Okun, 2004). Bologna and Appel's and Nanni's methods again use the same features and the ensemble-based classifiers (Bologna and Appel, 2002; Nanni, 2006).

Finally, method by Shen and Chou uses a new representation that includes 283 features and the ensemble-based classifier. They substituted composition vector from the feature set proposed by Ding and Dubchak with 178 features that implement pseudo-amino acid composition. When compared with the competing methods, PFRES uses only 36 features, which is 70% less features than the representation applied in Ding and Dubchak (2001), Bologna and Appel (2002), Nanni (2006) and Okun (2004), and nearly 90% less features than the representation proposed in Shen and Chou (2006). Table 3 shows that PFRES provides 6.3–12.4% higher accuracy than the prior methods. When compared with the best performing competing method by Shen and Chou, prediction with PFRES results in substantial $6.3/37.9 = 17\%$ error rate reduction. PFRES provides superior accuracy for 13 out of 27 folds, while method by Shen and Chou provides the best predictions for nine folds.

The statistical significance of the differences between accuracies obtained by the proposed and the competing methods over the 27 proteins folds was investigated using paired *t*-test. The corresponding *t*-values for the differences between PFRES and PFP (Shen and Chou, 2006), SE (Nanni, 2006), DIMLP (Bologna and Appel, 2002), HKNN (Okun, 2004) and SVM (Ding and Dubchak, 2001) methods equal 2.44, 3.18, 2.82, 3.12 and 4.12, respectively. As the critical *t*-value for the standard 0.05 significance level equals 1.71, the test shows that the proposed method provides statistically significantly better predictions than the predictions of the five competing methods. We also note that critical *t*-values for stronger, 0.01 and 0.005, significance levels equal 2.48 and 2.78, respectively.

3.5 Impact of the quality of the secondary structure predicted by PSI-PRED

Since 15 features proposed in this article were generated from the secondary structure predicted by PSI-PRED, we further analyze the impact of the quality of the predicted secondary structure on the accuracy of the fold classification. For the test set 2, the average accuracy of the predicted secondary structure was 75.4%. We divided the test set 2 into two subsets with sequences for which the secondary structure was predicted with accuracy below and above the average, correspondingly. The PFRES was evaluated on each of these subsets independently, see Table 4. The prediction accuracy for the second subset was 67.3%, while for the first subset it was slightly lower and equal 65.2%. As expected, higher quality of predicted secondary structure results in higher accuracy of fold classification. At the same time, this difference is relatively small, i.e. 2%, while the difference in accuracy of the predicted secondary structure between these two subsets was much larger (over 13%, see Table 4). This shows that the proposed method provides relatively stable quality of predictions with respect to the quality of the predicted secondary structure. We also note that current secondary structure prediction methods achieve the average accuracy close to 80%, e.g. EVA server reports that PSI-PRED provides the average accuracy 77.9% for 224 proteins (tested between April 2001 and September 2005), and Porter provides the average accuracy of 79.8% for 77 proteins (February 2005 to March 2006) (Eyrich *et al*., 2001). Since the average accuracy of the predicted secondary structure for sequences in the test set 2 was 75.4%, we believe that the presented test results provide a reliable estimate of the future performance of the proposed method.

Table 4. Average accuracy of predicted secondary structure and accuracy of fold classification for two subsets of test set 2; subset 1 includes sequences for which secondary structure was predicted with accuracy below 75.4%; subset 2 includes the remaining sequences

	Number of sequences	Average accuracy of predicted secondary structure (%)	Accuracy of fold classification with PFRES (%)
Subset1	379	67.6	65.2
Subset2	529	81.1	67.3
Total	908	75.4	66.4

4 CONCLUSIONS

A high quality predictor for the protein fold classification would be beneficial for *in silico* prediction of tertiary structure of proteins with low sequence identity, since it would allow for the determination of structural similarity without the sequence similarity. To this end, we propose PFRES method that uses a novel protein sequence representation, which consists of a small set of 36 features, and applies a carefully designed ensemble classifier. The proposed feature representation that is utilized by PFRES includes PSI-BLAST profile-based composition vector, features based on secondary structure predicted with PSI-PRED and sequence length. The experimental evaluation of the proposed fold classification method was performed with a standard benchmark dataset and another large set of over 900 sequences, both with chains with identity below 35% with respect to the training sequences. Using the benchmark set, PFRES is shown to predict the protein folds with 68.4% accuracy, which is over 6% higher than the accuracy of the best existing method. The results also show that the fold classification accuracy of the proposed method is statistically significantly better than the accuracy of all competing methods. Similar performance, i.e. 66.4% was achieved by the proposed method on the second test set. At the same time, PFRES uses 70–90% less features to represent sequences when compared with the existing methods. The proposed PSI-BLAST profile-based composition vector, which imbeds evolutionary information, was compared with commonly used sequence-based composition vector. Our empirical tests with six machine-learning classifiers have shown that the PSI-BLAST profile-based composition vector is superior to the composition vector. The new representation can be extended to other protein prediction tasks that currently apply AA composition, e.g. prediction of structural class, secondary structure content, membrane protein type, enzyme family, etc. to improve their accuracy.

ACKNOWLEDGEMENTS

K.C.'s research was supported by the Alberta Ingenuity Scholarship and NSERC Canada. L.K. acknowledges support from NSERC Canada.

Conflict of Interest: none declared.

REFERENCES

Aha,D. and Kibler,D. (1991) Instance-based learning algorithms. *Mach. Learn.*, **6**, 37–66.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **17**, 3389–3402.

Andreeva,A. *et al.* (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.

Birzele,F. and Kramer,S. (2006) A new representation for protein secondary structure prediction based on frequent patterns. *Bioinformatics*, **22**, 2628–2634.

Bologna,G. and Appel,R.D. (2002) A comparison study on protein fold recognition. In *Proceedings of the 9th International Conference on Neural Information Processing*. Vol. 5, pp. 2492–2496.

Bujnicki,J.M. (2006) Protein structure prediction by recombination of fragments. *Chem. BioChem.*, **7**, 19–27.

Chandonia,J.M. and Brenner,S.E. (2006) The impact of structural genomics: expectations and outcomes. *Science*, **311**, 347–351.

Chen,K. *et al.* (2007) Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. *BMC Struct. Biol.*, **7**, 25.

Chou,K.C. (2005) Progress in protein structural class prediction and its impact to bioinformatics and proteomics. *Curr. Protein Pept. Sci.*, **6**, 423–436.

Chothia,C. (1992) Proteins. One thousand families for the molecular biologist. *Nature*, **357**, 543–544.

Cleary,J.G. and Trigg,L.E. (1995) K*: an instance-based learner using an entropic distance measure. In *Proceedings of the 12th International Conference on Machine Learning*, pp. 108–114.

Ding,C.H. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.

Eyrich,V.A. *et al.* (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, **17**, 1242–1243.

Holley,L.H. and Karplus,M. (1989) Protein secondary structure prediction with a neural network. *Proc. Natl Acad. Sci. USA*, **86**, 152–156.

John,G.H. and Langley,P. (1995) Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pp. 338–345.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Jones,D.T. (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, **23**, 538–544.

Kerthi,S.S. *et al.* (2001) Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput.*, **13**, 637–649.

Kim,H. and Park,H. (2004) Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins*, **54**, 557–562.

Kurgan,L and Chen,K. (2007) Prediction of protein structural class for the twilight zone sequences. *Biochem. Biophys. Res. Commun.*, **357**, 453–460.

Le,C.S. and Houwelingen,J.C. (1992) Ridge estimators in logistic regression. *Appl. Stat.*, **41**, 191–201.

Leo,B. (2001) Random forests. *Mach. Learn.*, **1**, 5–32.

Levitt,M. (2007) Growth of novel protein structural data. *Proc. Natl Acad. Sci. USA*, **104**, 3183–3188.

Levitt,M. and Chothia,C. (1976) Structural patterns in globular proteins. *Nature*, **261**, 552–558.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lin,K. *et al.* (2005) A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics*, **21**, 152–159.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–5340.

Nanni,L. (2006) A novel ensemble of classifiers for protein fold recognition. *Neurocomputing.*, **69**, 2434–2437.

Okun,O. (2004) Protein fold recognition with K-local hyperplane distance nearest neighbor algorithm In *Proceedings of the 2nd European Workshop on Data Mining and Text Mining in Bioinformatics*. Vol. 1, pp. 51–57.

Paiardini,A. *et al.* (2004) Evolutionarily conserved regions and hydrophobic contacts at the superfamily level: the case of the fold-type I, pyridoxal-5′-phosphate-dependent enzymes. *Protein Sci.*, **13**, 2992–3005.

Quian,N. and Sejnowski,T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, **202**, 865–884.

Reinhardt,A. and Eisenberg,D. (2004) DPANN: improved sequence to structure alignments following fold recognition. *Proteins*, **56**, 528–538.

Ruan,J. *et al.* (2006) Quantitative analysis of the conservation of the tertiary structure of protein segments. *Protein J.*, **25**, 301–315.

Seewald,A.K. (2002) How to make stacking better and faster while also taking care of an unknown weakness. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 554–561.

Seewald,A.K. and Fuernkranz,J. (2001) An evaluation of grading classifiers. In *Proceedings of 4th International Conference on Advances in Intelligent Data Analysis*, pp. 115–124.

Shen,H.B. and Chou,K.C. (2006) Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, **22**, 1717–1722.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Tomii,K. *et al.* (2005) Protein structure prediction using a variety of profile libraries and 3D verification. *Proteins*, **61** (7), 114–121.

Tress,M. *et al.* (2005) Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins*, **61** (7), 27–45.

Wang,G. *et al.* (2005) Assessment of fold recognition predictions in CASP6. *Proteins*, **61 (Suppl. 7)**, 46–66.

Witten,I. and Frank,E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.

Yu,L. and Liu,H. (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In *Proceedings of the 10th International Conference on Machine Learning*, pp. 856–863.

Yu,Y.K. *et al.* (2006) Retrieval accuracy, statistical significance and compositional similarity in protein sequence database searches. *Nucleic Acids Res.*, **34**, 5966–5973.

Zhang,Y. and Skolnick,J. (2005) The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl Acad. Sci. USA*, **102**, 1029–1034.