Accurate prediction of protein structural classes using functional domains and predicted secondary structure sequences

Amin Ahmadi Adl, Abbas Nowzari-Dalini, Bin Xue, Vladimir N. Uversky, & Xiaoning Qian

Department of Computer Science & Engineering, University of South Florida, Tampa, FL, 33620, USA
Center of Excellence in Bioinformatics, School of Mathematics, Statistics, and Computer Science, University of Tehran, Tehran, Iran
Department of Molecular Medicine, College of Medicine, University of South Florida, Tampa, FL, 33612, USA
Institute for Biological Instrumentation, Russian Academy of Science, 142290, Pushchino, Moscow Region, Russia

Published online: 18 Apr 2012.

To cite this article: Amin Ahmadi Adl, Abbas Nowzari-Dalini, Bin Xue, Vladimir N. Uversky & Xiaoning Qian (2012) Accurate prediction of protein structural classes using functional domains and predicted secondary structure sequences, Journal of Biomolecular Structure and Dynamics, 29:6, 1127-1137, DOI: 10.1080/07391102.2011.672626

To link to this article: http://dx.doi.org/10.1080/07391102.2011.672626

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions
Accurate prediction of protein structural classes using functional domains and predicted secondary structure sequences

Amin Ahmadi Adl, Abbas Nowzari-Dalini, Bin Xue, Vladimir N. Uversky and Xiaoning Qian

“Department of Computer Science & Engineering, University of South Florida, Tampa, FL 33620, USA; “Center of Excellence in Bioinformatics, School of Mathematics, Statistics, and Computer Science, University of Tehran, Tehran, Iran; “Department of Molecular Medicine, College of Medicine, University of South Florida, Tampa, FL 33612, USA; “Institute for Biological Instrumentation, Russian Academy of Science, 142290, Pushchino, Moscow Region, Russia

Communicated by Ramaswamy H. Sarma

(Received 17 September 2011; final version received 12 December 2011)

Protein structural class prediction is one of the challenging problems in bioinformatics. Previous methods directly based on the similarity of amino acid (AA) sequences have been shown to be insufficient for low-similarity protein data-sets. To improve the prediction accuracy for such low-similarity proteins, different methods have been recently proposed that explore the novel feature sets based on predicted secondary structure propensities. In this paper, we focus on protein structural class prediction using combinations of the novel features including secondary structure propensities as well as functional domain (FD) features extracted from the InterPro signature database. Our comprehensive experimental results based on several benchmark data-sets have shown that the integration of new FD features substantially improves the accuracy of structural class prediction for low-similarity proteins as they capture meaningful relationships among AA residues that are far away in protein sequence. The proposed prediction method has also been tested to predict structural classes for partially disordered proteins with the reasonable prediction accuracy, which is a more difficult problem comparing to structural class prediction for commonly used benchmark data-sets and has never been done before to the best of our knowledge. In addition, to avoid overfitting with a large number of features, feature selection is applied to select discriminating features that contribute to achieve high prediction accuracy. The selected features have been shown to achieve stable prediction performance across different benchmark data-sets.

Keywords: Protein structural class prediction; functional domains; predicted secondary structure sequences; protein secondary structure propensity; disordered proteins; support vector machines (SVMs); feature selection

Introduction

Functionalities of proteins have been commonly believed to be determined by their unique 3D (dimensional) structures (Chou, 2006), which are determined by the exact spatial position of each atom. However, for simplicity, proteins typically are first classified into several structural folding classes, based on the type, amount, and spatial arrangement of their amino acid (AA) residues into potential secondary structure elements. For example, in structural classification of proteins (SCOP) (Murzin, Brenner, Hubbard, & Chot, 1995), proteins are annotated by structural class labels as the first step for their 3D structure annotations, among which there are four major structural classes denoted as α, β, αβ, and α + β. These four major classes cover 82, 89, and 84% of protein folds, families, and super-families in SCOP. Proteins in the class α have α-helices as the dominant secondary structure. Similarly, secondary structures of proteins in the class β are mostly dominated by β-strands. In the αβ and α + β classes, there are significant amounts of both α-helices and β-strands. In αβ, β-strands create parallel β-sheets; while in α + β class, β-strands create anti-parallel β-sheets (Murzin et al., 1995).

In addition to proteins with well-defined 3D structure, the protein universe includes intrinsically disordered proteins (IDPs) and proteins with intrinsically disordered regions (IDRs). These IDPs and IDRs are biologically active and yet fail to form specific 3D structure, existing instead as collapsed or extended dynamically mobile conformational ensembles (Daughdrill, Pielak, Uversky, Cortese, & Dunker, 2005; Dunker et al., 1998; Dunker et al., 2001; Tompa, 2002; Uversky & Dunker, 2010; Uversky, Gillespie, & Fink, 2000; Wright & Dyson, 1999). These proteins are highly abundant in any given proteome (Dunker et al., 2001; Dunker, Obradovic, Romero, Garner, & Brown, 2000; Uversky, 2010; Uversky & Dunker, 2010; Ward, Sodhi, McGuffin, Buxton, & Jones, 2004), and intrinsic disorder is shown to be crucial for determining protein functionality. Biological activities of disordered proteins (which are typically involved in

*Corresponding author. Email: xqian@cse.usf.edu

ISSN 0739-1102 print/ISSN 1538-0254 online
Copyright © 2012 Taylor & Francis
http://dx.doi.org/10.1080/07391102.2011.672626
http://www.tandfonline.com
regulation, signaling and control pathways (Dunker, Cortes, Romero, Lakoucheva, & Uversky, 2005; Lakoucheva, Brown, Lawson, Obradovic, & Dunker, 2002; Uversky, Oldfield, & Dunker, 2005) complement the functional repertoire of ordered proteins that have evolved mainly to carry out efficient catalysis (Radivojac et al., 2007; Vucetic et al., 2007; Xie, Vucetic, Lakoucheva, Oldfield, Dunker, Obradovic et al., 2007; Xie, Vucetic, Lakoucheva, Oldfield, Dunker, Uversky et al., 2007). IDPs are known to be associated with various human diseases, such as cancer, cardiovascular disease, amyloidosis, and neurodegenerative diseases (Uversky, Oldfield, & Dunker, 2008). IDPs and IDRs are very different from ordered proteins and domains in their AA sequences, and these differences were used to develop various predictors of intrinsic disorder (Dosztanyi & Tompa, 2008; Ferron, Longhi, Canard, & Karlin, 2006; He et al., 2009). These proteins and regions possess very wide structural diversity; and both extended (random coil-like) regions with perhaps some secondary structure and collapsed (partially folded or molten globule-like and pre-molten globule-like) domains with poorly packed side chains are considered as intrinsically disordered (Dunker et al., 2001; Uversky, 2002, 2011). Despite the recent progress and success in protein structural class prediction, prediction for IDPs and IDRs has not been studied in the past.

Due to the rapid growth in the number of discovered sequences and the burden of experimental screening methods regarding time and cost to find the 3D structure and thereafter functionality of the proteins, it will help to speed up and reduce the cost for protein annotation if one can develop fast computational methods that use all the available information to predict at least some important characteristics of protein structure to screen large protein data-sets and focus the limited resources on proteins of interest. As stated in Mizianty and Kurgan (2009), protein structural class prediction has a wide range of potential applications, including prediction of protein (un)folding rates, prediction of DNA binding sites, reduction of conformation search space, and implementation of heuristic approaches to predict tertiary structure. In this paper, we focus on this challenging problem of determining protein structural classes using novel features derived from protein AA sequences, especially for low-similarity as well as partially disordered protein data-sets. The latter has not been studied in the literature.

Related work

Computational protein structural class prediction has been a challenging task for bioinformaticians, whose aim is to find a prediction model to automatically determine the structural class based on the protein AA sequence (Chen, Kurgan, & Ruan, 2008; Chou, 1995; Chou & Zhang, 1995; Ding, Zhang, & Chou, 2007; Luo, Feng, & Liu, 2002; Sun & Huang, 2006). Typically, AA sequences with variable lengths are first analyzed to derive fixed length feature vectors, which are then used by machine learning algorithms to build a prediction model.

Different features and learning algorithms have been implemented to solve this problem (Chen et al., 2008; Chou, 1995; Chou & Zhang, 1995; Ding et al., 2007; Kurgan et al., 2008; Luo et al., 2002; Sun & Huang, 2006; Yang et al., 2009). Many previous protein structural class prediction methods are based on simple sets of sequence-based features, which are directly computed from AA sequences such as, for example, the frequency of each AA in given proteins. These simple features typically ignore the sequential order of AAs and the relationships between distant AAs. Here, the terms distant AAs or distance residues refer to the residues that are far away in the protein AA sequences. Often, the corresponding prediction methods have mediocre performance (Chou, 1995; Chou & Zhang, 1995). Recently, high-order sequence-based features that take into account the order of AAs and the relationships between distant AAs, such as composition of short polypeptides (Luo et al., 2002; Sun & Huang, 2006), pseudo AA composition (Ding et al., 2007), collocation of AA, and position-specific scoring matrix profiles computed by position-specific iterative basic local alignment search tool (PSI Blast) (Chen et al., 2008), have been explored to improve the performance. Since these advanced features are sequence dependent, and since it has been conjectured that similar sequences share similar folding patterns, these new tools have had success for the analysis of high-similarity proteins. However, the performance degrades substantially for low-similarity protein data-sets, in which AA sequences share similarities below 40% (Yang & Chen, 2010).

In order to achieve high prediction accuracy for low-similarity proteins, several new feature sets have been proposed, including those based on predicted secondary structure propensities (Kurgan et al., 2008; Mizianty & Kurgan, 2009; Yang et al., 2010). They exploit the fact that proteins with low sequence similarity but in the same structural class are likely to have high similarity in their corresponding secondary structure elements. Novel computational tools that utilized these features based on predicted secondary structure propensities have achieved highest accuracy described in literature so far, between 80 and 83% on several low-similarity benchmark data-sets (Kurgan et al., 2008; Mizianty & Kurgan, 2009; Yang et al., 2010).

Despite the recent success of these secondary structure based features, they still mostly use local properties of protein residues, and therefore may not capture the useful relationships among distant AAs. Another set of features called functional domain (FD) composition has been proposed to further improve the prediction performance (Chou & Cai, 2004). The FD information has also been used (either independently or combined with other information) in other computational predicting problems such as predicting protein subcellular location (Chou & Cai, 2002). FD is the core of a protein that plays the major role for its function. That is why in determining the 3D structure of a protein by experiments or by computational modeling, the first priority has always been focused on its FD. These FD features can be extracted
by analyzing sequences using information available in the InterPro database (Hunter et al., 2009), in which a large number of protein sequence signatures, motifs, and domains with potential functional sites have been annotated and integrated from multiple databases, including Gene3D (Yeats et al., 2008), PANTHER (Mi, Guo, Kejariwal, & Thomas, 2007), Pfam (Finn et al., 2008), PIRSF (Nikolskaya, Arighi, Huang, Barker, & Wu, 2006), PRINTS (Attwood et al., 2003), ProDom (Bru et al., 2005), PROSITE (Hulo et al., 2006), SMART (Letunic et al., 2006), SUPERFAMILY (Wilson, Madera, Vogel, Chothia, & Gough, 2007), and TIGRFAMs (Haft, Selengut, & White, 2003). These FD features are assumed to be more sensitive to the order of AA residues and correlations between distance residues than previously derived features, as they capture more global information by relating to the longer subsequences. They also can capture the functional similarity between proteins that may be useful in structural prediction for low-similarity data-sets. Preliminary results in Chou and Cai (2004) have been reported with more than 90% accuracy based on a low-similarity database designed by the authors for structural prediction into seven classes. However, a comprehensive validation of this new set of FD features for protein structural class prediction on widely recognized benchmark data-sets is still missing as yet.

**Contributions**

The main goal of our research is to explore potential ways to integrate the newly derived features, including predicted secondary structure propensities and FD features, to improve the performance of a predictor of protein structural classes in low-similarity data-sets. In particular, we focus our investigation on the importance of FD features in protein structural class prediction and on performing comprehensive experiments to show that these features are complementary to secondary structure based features and contribute to the substantial increase of the accuracy in protein structural class prediction. The major contributions of this paper include the following:

1. We have implemented a multi-class support vector machine (SVM) classifier to predict protein structural class using potential combinations of different feature sets, including predicted secondary structure propensities and the FD features.

2. We have evaluated the prediction performance on five different benchmark data-sets. The results demonstrate that the integration of FD features substantially increases the prediction accuracy ranging from 3 to 6% improvement for different data-sets. The proposed structural class prediction method has achieved the highest accuracies on three common benchmark data-sets among the existing state-of-the-art methods. We note that the number of structurally annotated proteins is rather limited. For example, only around 38,221 PDB entries with 110,800 domains or proteins have known structural class labels in SCOP (as of February 2009), while there are more than 8,000,000 nonredundant protein sequences in the Protein database at the National Center for Biotechnology Information (NCBI). Hence, 1% improvement in accuracy can help in finding the accurate structural class labels for about 80,000 proteins.

3. In addition to low-similarity benchmark data-sets, we evaluate the structural class prediction method using new benchmark data-sets of proteins with different disorder levels, which is studied for the first time for protein structural class prediction to the best of our knowledge.

4. Correlation-Based Feature Selection (CFS) (Hall, 1999) has been implemented to identify a small number of uncorrelated features to avoid overfitting, but still to achieve high prediction accuracy. The experimental results across different benchmark data-sets have shown that identified small feature sets are stable and lead to consistently high prediction accuracy when using selected features from different training data-sets.

**Materials and methods**

In this paper, we propose to implement a multi-class SVM classifier with feature selection for protein structural class prediction using three recently proposed feature sets. We discuss the technical details in this section.

**Features**

Among features exploited in literature, we have chosen three recently proposed feature sets, as they have been reported to obtain high prediction accuracy. The first set is FD features extracted from the InterPro data-set using a method similar to Chou and Cai (2004). The other two sets are predicted secondary structure elements by methods given in Kurgan and Chen (2008) and Yang et al., 2010 (SS1 and SS2). We note that we did not include the simple sequenced-based features, since it has been shown that these features did not perform well on data-sets with low sequence similarity. For example, in Kurgan et al. (2008), 2146 features extracted from AA sequences were studied together with 176 features from secondary structure elements. Among those 2146 sequence-based features, only one has been selected with marginally improved accuracy by 0.4%.

**FD feature set**

First, we will discuss a set of FD features that can be obtained based on the InterPro database (http://www.ebi.ac.uk/interpro/). These features were first introduced in Chou and Cai (2004) for protein structural class prediction. InterPro Release 30.0 (1 December 2010) contains 21,178 entries, representing sequence signatures corresponding to possible FDs. With that, any protein can be represented as a 21,178 dimensional binary vector \( f = (f_1, \ldots, f_{21178})^T \), in which \( f_i = 1 \) denotes that the pro-
tein has a corresponding FD and zero otherwise. Using InterPro Scan (IPRSCAN) (Zdobnov & Apweiler, 2001), we can search the InterPro database for any given protein and obtain its binary feature vector based on the matched sequence signatures in the database.

**Secondary structure based feature sets (SS1 and SS2)**

Based on the AA sequence of a given protein, the corresponding sequence of secondary structure elements can be predicted as shown in Figure 1, where each residue can belong to a $\alpha$-helix (H), $\beta$-sheet (E), or a random coil (C). In order to extract secondary structure based features, protein structure prediction server (PSI-PRED) (Mcguffin, Bryson, & Jones, 2000) has been first applied to predict secondary structure propensities for proteins under study. Based on the predicted secondary structure elements, different features can be derived for protein structural class prediction. In this paper, we focus on two feature sets proposed by two state-of-the-art structural class prediction methods described in Kurgan et al. (2008) and Yang et al. (2010).

The first secondary structure feature set (SS1) was proposed in Kurgan et al. (2008). This set contains 86 features extracted from the predicted secondary structure elements, which include frequencies and composition moments of secondary structure elements, as well as the count and the length of segments of secondary structure elements. Detailed information about SS1 can be found in Kurgan et al. (2008) and in Text S1. The second feature set (SS2) was derived in Yang et al. (2010). This set contains 24 features based on the recurrence quantification (Marwan, Romano, Thiel, & Kurths, 2007), K-string entropy, and segment-based analysis of predicted secondary structure sequences. For more information about SS2, one can refer to Yang et al. (2009) and Text S1.

**Classifier**

We apply a four class SVM for our protein structural class prediction model as we use four folding class labels introduced earlier. We note that the proposed method can be extended for seven class prediction in a straightforward manner when necessary. The current setup is for the convenience to compare with other existing methods.

In order to use SVM for our problem, six standard binary SVM classifiers, namely, $\alpha$ vs. $\beta$, $\alpha$ vs. $\alpha + \beta$, $\beta$ vs. $\alpha + \beta$, and $\alpha + \beta$ vs. $\alpha + \beta$, have been trained. The predictions from these classifiers finally have been combined to determine the final class labels using a pairwise coupling method as presented in Hastie and Tibshirani (1998). Specifically, for a new protein with feature vector $\mathbf{f}$, the following conditional probability can be computed based on any binary classifier:

$$P(\mathbf{f} \in A \mid \mathbf{f} \in A \text{ or } \mathbf{f} \in B) = \frac{P(\mathbf{f} \in A)}{P(\mathbf{f} \in A) + P(\mathbf{f} \in B)},$$

where $A$ and $B$ are the corresponding class labels for a binary classifier. Based on training results of each binary SVM classifier, $P(\mathbf{f} \in A \mid \mathbf{f} \in A \text{ or } \mathbf{f} \in B)$ can be estimated by a logistic model as suggested by Platt (2000). With these conditional probabilities, we solve the linear equations based on Chou (2006) to compute $P(\mathbf{f} \in \alpha)$, $P(\mathbf{f} \in \beta)$, $P(\mathbf{f} \in \alpha + \beta)$, and $P(\mathbf{f} \in \alpha + \beta + \gamma)$. As there are four variables with six equations, the solution may not exist in general. To solve this problem, pairwise coupling (Hastie & Tibshirani, 1998) is implemented to estimate the solution. The final class label is determined based on the highest probability within these four probabilities. Figure 2 illustrates the multi-class SVM classifier.

We implement gaussian kernel SVM with kernel size parameter $g$ and complexity parameter $C$. We search for best performing parameters for SVM by grid search during training. This classifier is implemented using Wai-kato environment for knowledge analysis (WEKA) (Witten & Frank, 2005) machine learning tool libraries. Further implementation details can be found in Text S1.

**Feature selection**

Due to the integration of FD features, the dimensionality of features in our learning algorithm is extremely high. To avoid overfitting, we adopt a CFS method (Hall, 1999) to reduce the feature dimension. CFS is a filtering method which identifies a small set of nonredundant features that are highly correlated with the outcome while having low correlation among themselves. Specifically, we search through the feature space for a good feature subset $S$, in which we would like to maximize the merit of $S$ with $K$ features:

$$\text{Merits}_S = \frac{K\bar{r}_{\mathbf{f}}}{\sqrt{K + K(K-1)\bar{r}_{\mathbf{f}}}},$$

in which $\bar{r}_{\mathbf{f}}$ measures the average dependence between $K$ features $\mathbf{f}$’s and the class label $c$, and $\bar{r}_{\mathbf{f}}$ measures the average dependence among $K$ features in $S$. The value of $\text{Merits}_S$ increases when the selected features are highly informative about the outcome, but decreases when there is a high correlation among those features. We have implemented hill-climbing optimization as in the best-first search (Pearl, 1984) with five levels of back-tracking, which iteratively expands the feature subset $S$ starting from an empty set to identify a better $S$ based on the merit value among all the possible expansions at each step until there are five consecutive nonimproving expansions.

Protein: GKGDPPKPRGKMSSAFFFVCQTSREEKKHHHPDASVNFSEFSKCSERWKTMASAEKFKEDMAKADKARYEMKTYIPPG

SS: CCCCCCCCCCCCCCHHHHEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

Figure 1. Predicted secondary structure elements.
We adopt the WEKA (Witten & Frank, 2005) machine learning package for CFS. Continuous features are first discretized to nominal values (Fayyad & Irani, 1993). The above average dependences $r_{cf}$ and $r_{df}$ are computed by information gain. For example, given a feature $f$ and the outcome $c$, the information gain about $c$ after observing $f$ can be computed as the following:

$$\text{Gain}(c|f) = H(c) - H(c|f), \quad (3)$$

Where

$$H(c) = -\sum_c P(c) \log_2(P(c)) \quad (4)$$

is the entropy, and

$$H(c|f) = -\sum_f P(f) \sum_c P(c|f) \log_2(P(c|f)) \quad (5)$$

is the conditional entropy.

**Results and discussion**

In our experiments, we comprehensively evaluate the structural class prediction performance on a few benchmark data-sets using the previously introduced three sets of features, namely, SS1, SS2, and FD. The main goal is to validate the hypothesis:

Functional domain (FD) features capture the useful relationships among distant AAs that are critical for protein folding. The integration of FD features together with secondary structure based feature sets (SS1 and/or SS2) can substantially increase the protein structural class prediction accuracy, especially for low-similarity proteins as well as partially disordered proteins.

**Performance comparison for low-similarity proteins**

First, in order to demonstrate that the integration of FD features can substantially improve the performance, we compare the structural class prediction accuracy of our method with the performances of the two state-of-the-art methods presented in Kurgan et al. (2008) and Yang et al. (2010) using different possible combinations of three feature sets.

**Data-sets**

The proposed method is tested on three low-similarity protein data-sets that are widely used in the literature (Kurgan et al., 2008; Mizianty & Kurgan, 2009; Yang et al., 2010). The first two data-sets, referred to as 25PDB and 1189, respectively, are downloaded from RCSB Protein Data Bank (www.pdb.org) Berman, 2000 with the PDB IDs listed in the paper (Kurgan & Homaeian, 2006). The data-set 25PDB contains 1673 proteins with the pairwise sequence identity being about 25%, whereas the data-set 1189 contains 1092 proteins with 40% sequence identity. The third protein data-set, referred to as 640, was first studied in Chen et al. (2008). It contains 640 proteins with 25% sequence identity. There are 76 protein sequences that overlap among three data-sets. The numbers of common sequences between each pair of data-sets are 357 (for 640 and 1189), 78 (for 640 and 25PDB), and 205 (for 1189 and 25PDB), respectively. The AA sequences in these data-sets represent protein domains rather than the complete protein AA sequences. Protein structural classification labels are retrieved from the database SCOP (Murzin et al., 1995).

As we explained earlier, SS1 contains 86 features and SS2 has 24 features. For the FD features, we remove the FDs which do not appear in any of the proteins in a given data-set. After removing these features, we have 2400 FD features for 25PDB data-set, 1648 FD features for 1189 data-set, and 1371 FD features for 640 data-set.

**Performance comparison**

In our experiments, the parameters of SVMs are tuned by the grid search method based on 10-fold cross-
validation considering its computational efficiency. Among the independent data-set tests, sub-sampling (e.g., 5 or 10-fold cross-validation) test, and Jackknife test (Efron, 1982), which are often used for examining the accuracy of a statistical prediction method (Chou & Zhang, 1995), the Jackknife test has been considered more rigorous and the least arbitrary that can always yield a unique result for a given benchmark data-set. Thus, in order to faithfully compare our method with other state-of-the-art methods, we also estimate the prediction accuracy using the Jackknife method (Efron, 1982). The detailed introduction can be found in Text S1.

In Figure 3, the accuracies based on 10-fold cross-validation for 640, 1189, and 25PDB data-sets are plotted for seven different combinations of features, i.e., FD, SS1, SS2, SS1 + SS2, FD + SS1, FD + SS2, and FD + SS1 + SS2. In the first set of experiments, we have used the complete set of features without using any feature selection algorithm. More detailed information about the results are provided in Tables 1–3 in Text S1. From these results, it is clear that combining FD with SS1 and/or SS2 features can substantially improve the accuracy of protein structural class prediction. From the figure, we have the highest accuracies of 86.09, 90.01, and 88.22% for 640, 1189, and 25PDB data-sets, respectively. The highest accuracies for 1189 and 25PDB are obtained with the FD + SS2 combination, and the highest accuracy for 640 data-set is obtained using the FD + SS1 + SS2 combination.

Table 1. The number of selected features.

| Data-set | Total number of features (FD + SS1 + SS2) | Selected features using CFS (FD + SS1 + SS2) |
|----------|------------------------------------------|---------------------------------------------|
| 640      | 1481(1371 + 86 + 24)                     | 32(8 + 15 + 9)                              |
| 1189     | 1758(1648 + 86 + 24)                     | 50(21 + 23 + 6)                             |
| 25PDB    | 2510(2400 + 86 + 24)                     | 57(22 + 26 + 9)                             |

We further test the prediction performance with feature selection. In this set of experiments, we implement the CFS approach to reduce the number of features in all seven different combinations. Table 1 gives the number of selected features for the FD + SS1 + SS2 combination for three data-sets, respectively. More details are provided in Text S1. The accuracies obtained for the reduced sets of features are given in Figure 4. The highest accuracies are 85.15, 86.63, and 86.67% for 640, 1189, and 25PDB data-sets, respectively. All of them are obtained using the feature combination of all three feature sets: FD + SS1 + SS2. Although these accuracies are lower than the accuracies obtained using the complete feature sets, they are still noticeably higher than the accuracies of any previously proposed protein structural class prediction methods. Furthermore, after using CFS, we have a much smaller number of features. More detailed information about the results are provided in Tables 1–3 in Text S1.

We study the stability of these selected features by focusing on the FD + SS1 + SS2 feature combination. Some of the selected features are consistently appearing across three data-sets. The overlap contains 17 features, nine from SS1, four from SS2, and four from FD. The union of selected features contains 80 features, 33 from SS1, 11 from SS2, and 36 from FD. The actual selected features for three data-sets can be found in Tables 4 and 5 in Text S1. Based on the features selected in the previous experiments, we have designed six more experiments to study feature stability by evaluating the performance of selected features from each data-set across the two other data-sets. In other words, we have used the features...
selected for each data-set to train the prediction model for the two other data-sets and estimated the accuracies based on 10-fold cross-validation. The results are shown in Table 2, which clearly shows that the features identified independently for each data-set are stable and achieve comparable prediction accuracies across all three benchmark data-sets. Further detailed results can be found in Table 6 in Text S1.

To further assure that the integration of FD features improved the performance, we also have evaluated the prediction accuracy for different data-sets using the Jackknife method. The values of the SVM classifier parameters \((C, g)\) are based on the previous grid search for 10-fold cross-validation. Table 3 compares the overall accuracies obtained by our method using Jackknife evaluations and the highest accuracies using the other three state-of-the-art algorithms in Kurgan et al. (2008), Mizianty and Kurgan (2009), and Yang et al. (2010). These results further confirm the substantial improvement in the accuracy of the protein structural class prediction using the FD features.

### Performance comparison for fully structured and partially structured proteins

We have further tested the proposed method on two other data-sets with proteins with different levels of intrinsic disorder (Xue, Dunbrack, Williams, Dunker, & Uversky, 2010; Xue, Oldfield, Dunker, & Uversky, 2009) using combinations of FD and SS1 feature sets. Fully structured data-set (Structured) was selected from PDB by choosing only single-chain and non-membrane structures obtained by X-ray crystallography and characterized by unit cells with primitive space groups, and by removing structures that have ligands, disulfide bonds, and missing residues. The remaining sequences were clustered by using BLASTCLUST from NCBI (http://ncbi.nlm.nih.gov/BLAST) to regroup sequences with 25% and higher sequence identity into one cluster. The longest sequence of each cluster was chosen to compose the final data-set that has 554 chains with 113,895 residues. Partially structured data-set (Disordered) was selected from PDB X-ray struc-

---

**Table 2.** Performance of selected features across all data-sets.

| Testing data-set | Training data-set for feature selection |
|------------------|----------------------------------------|
|                  | 640 (%)  | 1189 (%) | 25PDB (%) |
| 640              | 85.15    | 85.93    | 85.62     |
| 1189             | 83.24    | 86.63    | 83.69     |
| 25PDB            | 85.23    | 85.29    | 86.67     |

**Table 3.** Performance comparison of our proposed method with the state-of-the-art methods for all three benchmark data-sets based on Jackknife.

| Data-set | (33) (SS1) | (26) custom(including SS1) | (35) (SS2) | This paper (FD + SS1 + SS2) | This paper (FD + SS1 + SS2 and CFS) |
|----------|------------|-----------------------------|------------|-----------------------------|-------------------------------------|
| 640      | N/A        | N/A                         | 83.10      | 86.25                       | 85.00                               |
| 1189     | N/A        | 83.50%                      | 81.30      | 89.65                       | 85.71                               |
| 25PDB    | 80.20%     | 81.40%                      | 82.90      | 87.80                       | 86.13                               |
tures with the resolution better than 3.0 angstroms, by choosing single-chain proteins with no ligands or partners. The sequences were clustered by using BLASTCLUST with a sequence identity cut-off of 30%. The longest sequence in each cluster was selected for inclusion. The resulting set of sequences was further filtered by keeping only those sequences that have 20 or more consecutive disordered residues as identified by xml2pdb (Dunbrack, 2010). Finally, there are 647 sequences with a total of 230,314 residues, within which 16,011 disordered residues are located within 1376 disordered regions. To evaluate the prediction performance of our method, we select sequences from these two data-sets that have structural class annotations in SCOP with one of the four major classes, which leads to the final 415 sequences in fully structured data-set and 332 sequences in partially structured data-set. Note that there is no overlap between these two new data-sets and the previous three data-sets.

The prediction results are shown in Figures 5 and 6. The first data-set contains low-similarity fully structured proteins, which resembles the properties of the previous three benchmark data-sets. As we expected, the integration of FD features again substantially improves the prediction accuracy for this data-set. The other data-set contains proteins with relatively high disorder level. As these proteins are naturally disordered (i.e. they contain regions that do not have stable tertiary and/or secondary structure), it is more difficult to predict their structural classes. Comparing to the previous four data-sets, the obtained accuracies are substantially lower. Nevertheless, the integration of FD features still improves the prediction accuracy, though the proteins in this data-set are partially structured.

We again have applied CFS to reduce the number of selected features to avoid overfitting. To study the stability of selected features across benchmark data-sets, we further evaluate the performance of selected features for structural class prediction for fully structured and partially structured proteins using four different sets of selected features from themselves as well as from the previous benchmark data-sets, including 640, and the union of 640, 1189, and 25PDB. Table 4 shows the results of this analysis. As expected, since the fully structured data-set shares high similarity with the previous three benchmark data-sets, the selected features from these data-sets perform comparably well on the fully

Figure 5. Accuracies using three different combinations of features for fully and partially structured data-sets.

Figure 6. Accuracies using three different combinations of features with CFS for fully and partially structured data-sets.
structured data-set. While the partially structured proteins contain regions with no stable structure and may have specific sequence characteristics, they are more difficult to predict. Hence, the corresponding prediction accuracy is relatively low and the selected features from the other data-sets perform relatively poorly in comparison with the performance of the selected features based on the partially structured data-set itself. We have also studied the selected features from these four different training data-sets. The detail can be found in Tables 6, 7, 8, 9, and 10 in Text S1. Using each individual data-sets, around 30 features are selected in each setting while they have large overlap with each other (over 30%). This suggests that the selected features are relatively stable and the integration of the FD features improves the prediction performance in general.

Conclusions

In this paper, we propose to explore both secondary structure propensity based features and FD features to improve the prediction accuracy of protein structural classes. We have performed thorough experimental comparison with the state-of-the-art structure class prediction algorithms using several benchmark data-sets. Our experimental results have demonstrated substantial improvement over the existing algorithms by integrating FD features. The experiments with feature selection have shown that we can achieve stable prediction performance using feature subsets selected across different benchmark data-sets. Finally, we have illustrated the proposed method has achieved reasonable performance for predicting partially structured proteins, which is an intrinsically more difficult task. The improvement introduced by FD features motivates many future applications in computational analysis of protein sequences, including the structure prediction as well as further analysis and understanding of IDPs using predicted FD features as they capture critical correlation information among distant AAs.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors, we shall make efforts in our future work to provide a web-server for the method presented in this paper.

Acknowledgments

The authors would like to thank Jian-Yi Yang and Dr. Lukasz Kurgan for making available the related data for three low-similarity benchmark data-sets.

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.org/10.1080/07391102.2011.672626.

References

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., ... Zygouri, C. (2003). Prints and its automatic supplement, preprints. Nucleic Acids Research, 31(2), 400–402.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., ... Bourne, P.E. (2000). The protein data bank. Nucleic Acids Research, 28, 235–242.

Bru, C., Courcelle, E., Carr’ere, S., Beausse, Y., Dalmar, S., & Kahn, D. (2005). The prodom database of protein domain families: More emphasis on 3D. Nucleic Acids Research, 33, D212–D215.

Chen, K., Kurgan, L.A., & Ruan, J. (2008). Prediction of protein structural class using novel evolutionary collocation-based sequence representation. Journal of Computational Chemistry, 29, 1596–1604.

Chou, K.C. (1995). Computer-Aided Drug Discovery. Proteins: Structure, Function, and Genetics, 21, 319–344.

Chou, K.C. (2006). Structural bioinformatics and its impact to biomedical science and drug discovery. Frontiers in Medicinal Chemistry, 3, 455–502.

Chou, K.C., & Cai, Y.D. (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. Journal of Biological Chemistry, 277, 45765–45769.

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

Chou, K.C., & Zhang, C.T. (1995). Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology, 30(4), 275–349.

Daughdrill, G.W., Pielak, G.J., Uversky, V.N., Cortese, M.S., & Dunker, A.K. (2005). Natively disordered proteins. In J. Buchner & T. Kieflhaber (Eds.), Handbook of protein folding (pp. 271–353). Weinheim: Wiley-VCH, Verlag GmbH & Co. KGaA.

Ding, Y.S., Zhang, T.L., & Chou, K.C. (2007). Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein and Peptide Letters, 14, 811–815.

Dosztanyi, Z., & Tompa, P. (2008). Prediction of protein disorder. Methods in Molecular Biology, 426, 103–115.
Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., & Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology*, 337, 635–645.

Wilson, D., Madera, M., Vogel, C., Chothia, C., & Gough, J. (2007). The SUPERFAMILY database in 2007: Families and functions. *Nucleic Acids Research*, 35, D308–D313.

Witten, I.H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Waltham, MA: Morgan Kaufmann.

Wright, P.E., & Dyson, H.J. (1999). Intrinsically unstructured proteins: Reassessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293, 321–331.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Obradovic, Z., & Uversky, V.N. (2007). Functional anthology of intrinsic disorder 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *Journal of Proteome Research*, 6, 1917–1932.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Uversky, V.N., & Obradovic, Z. (2007). Functional anthology of intrinsic disorder 1. Biological processes and functions of proteins with long disordered regions. *Journal of Proteome Research*, 6, 1882–1898.

Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, A.K., & Uversky, V.N. (2010). PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. *Biochimica et Biophysica Acta*, 1804(4), 996–1010.

Xue, B., Oldfield, C.J., Dunker, A.K., & Uversky, V.N. (2009). CDF it all: Consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. *FEBS Letters*, 583(9), 1469–1474.

Yang, J.Y., Peng, Z.L., & Chen, X. (2010). Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. *BMC Bioinformatics* S9(11).

Yang, J.Y., Peng, Z.L., Yu, Z.G., Zhang, R.J., Anh, V., & Wang, D. (2009). Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. *Journal of Theoretical Biology*, 257, 618–626.

Yeats, C., Lees, J., Reid, A., Kellam, P., Martin, N., Liu, X., & Orengo, C. (2008). Gene3D: Comprehensive structural and functional annotation of genomes. *Nucleic Acids Research*, 36, D414–D418.

Zdobnov, E.M., & Apweiler, R. (2001). InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9), 847–848.