Saudi Journal of Biological Sciences (2015) xxx, xxx–xxx

King Saud University

Saudi Journal of Biological Sciences

www.ksu.edu.sa
www.sciencedirect.com

The recognition of multi-class protein folds by adding average chemical shifts of secondary structure elements

Zhenxing Feng [a], Xiuzhen Hu [a,*], Zhuo Jiang [a], Hangyu Song [a], Muhammad Aqeel Ashraf [b]

[a] *Department of Sciences, Inner Mongolia University of Technology, Hohhot, China*
[b] *Water Research Unit, Faculty of Science and Natural Resources, University Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia*

Received 11 September 2015; revised 8 October 2015; accepted 12 October 2015

KEYWORDS

Multi-class protein folds;
The increment of diversity;
Average chemical shifts;
Secondary structure elements;
Secondary structure motifs;
Random Forest algorithm

Abstract The recognition of protein folds is an important step in the prediction of protein structure and function. Recently, an increasing number of researchers have sought to improve the methods for protein fold recognition. Following the construction of a dataset consisting of 27 protein fold classes by Ding and Dubchak in 2001, prediction algorithms, parameters and the construction of new datasets have improved for the prediction of protein folds. In this study, we reorganized a dataset consisting of 76-fold classes constructed by Liu et al. and used the values of the increment of diversity, average chemical shifts of secondary structure elements and secondary structure motifs as feature parameters in the recognition of multi-class protein folds. With the combined feature vector as the input parameter for the Random Forests algorithm and ensemble classification strategy, we propose a novel method to identify the 76 protein fold classes. The overall accuracy of the test dataset using an independent test was 66.69%; when the training and test sets were combined, with 5-fold cross-validation, the overall accuracy was 73.43%. This method was further used to predict the test dataset and the corresponding structural classification of the first 27-protein fold class dataset, resulting in overall accuracies of 79.66% and 93.40%, respectively. Moreover, when the training set and test sets were combined, the accuracy using 5-fold cross-validation was 81.21%. Additionally, this approach resulted in improved prediction results using the 27-protein fold class dataset constructed by Ding and Dubchak.

* Corresponding author.
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

1. Introduction

The large numbers of protein sequences generated in the post-genomic era has challenged researchers to develop a high-throughput computational method to structurally

http://dx.doi.org/10.1016/j.sjbs.2015.10.008
1319-562X © 2015 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

annotate these sequences. The protein fold reflects a key topological structure in proteins, as it contains three major aspects of protein structure: units of secondary structure, the relative arrangement of structures, and the overall relationship of protein peptide chains (Martin et al. 1998; Ming et al., 2015).

The proper spacial structure of a protein is highly correlated with its physiological functions. Abnormal protein folding may cause different diseases, for example, the neurodegenerative diseases such as Alzheimer's disease, spongiform encephalopathy, Parkinson's disease, mad cow disease etc. Thus, the correct identification of protein folds can be valuable for studies on pathogenic mechanisms and drug design (Thomas et al., 1995; Christopher and Michelle, 2004; Krishna and Grishin, 2005; Lindquist et al., 2001; Scheibel et al., 2004; Ma et al., 2002; Ma and Lindquist, 2002) and represents an important topic in bioinformatics.

In 2001, Ding and Dubchak (2001) constructed a dataset consisting of 27 protein fold classes using multiple feature parameters, including amino acid composition, predicted secondary structure, etc., and proposed support vector machines and neural network methods to predict the 27 protein fold classes, achieving an overall accuracy of 56.0%.

Subsequently, using the dataset constructed by Ding and Dubchak and identical feature parameters, several studies have suggested algorithmic improvements for protein fold identification. For example, Chinnasamy et al. (2005) introduced the phylogenetic tree and Bayes classifier for the identification of protein folds and achieved an overall accuracy of 58.2%. Nanni (2006) proposed a new ensemble of K-local hyperplanes based on random subspace and feature selection, achieving an overall accuracy of 61.1%. Guo and Gao (2008) presented a novel hierarchical ensemble classifier termed GAOEC (genetic-algorithm optimized ensemble classifier) and achieved an overall accuracy of 64.7%. Damoulas and Girolami (2008) proposed the kernel combination methodology for the prediction of protein folds and achieved an accuracy of 70%. Lin et al. (2013) exploited novel techniques to impressively increase the accuracy of protein fold classification.

Additional studies have suggested the selection of feature parameters to predict protein folds. For example, Shamim et al. (2007) used the structural properties of amino acid residues and amino acid residue pairs and achieved an overall accuracy of 65.2%. Dong et al. (2009) proposed a method termed ACCFold and achieved an overall accuracy of 70.1%. Nanni et al. (2010) proposed a method to extract features from the 3D structure and achieved significant improvement; however, this method does not solely rely on protein primary sequences to predict protein folds. Li et al. (2013) proposed a method termed PFP-RFSM and obtained improved results for protein fold identification.

Numerous studies have not only focused on the selection of feature parameters but also on the improvement of algorithms to identify protein folds. For example, Zhang et al. (2009) proposed an approach that utilizes the increment of diversity by selecting the pseudo amino acid composition, position weight matrix score, etc., and used these parameters to predict the 27 protein fold classes, with an overall accuracy of 61.1%. Shen and Chou (2006) applied the OET-KNN ensemble classifier to identify folds by introducing pseudo amino acids with sequential order information as a feature parameter and achieved an overall accuracy of 62.1%. Chen and Kurgan (2007) proposed the PFRES method using evolutionary information and predicted secondary structure, obtaining an accuracy of 68.4%. Ghanty and Pal (2009) proposed the fusion of heterogeneous classifiers approach, with features including the selected trio AACs and trio potential, and the overall recognition accuracy was 68.6%. Shen and Chou (2009) applied an identification method to protein folds using functional domain and sequential evolution information and achieved an overall accuracy of 70.5%. Yang and Kecman (2011) proposed a novel ensemble classifier termed MarFold, which combines three margin-based classifiers for protein fold recognition, and the overall prediction accuracy was 71.7%.

Additional studies have constructed and analyzed new 27-fold class datasets. For example, with a sequence identity less than 40%, Mohammad et al. (2007) constructed a dataset composed of 2554 proteins belonging to 27-fold classes, proposed structural properties of amino acid residues and amino acid residue pairs as parameters, and achieved an overall accuracy of 70.5% using 5-fold cross-validation. With sequence identity below 40%, Dong et al. (2009) constructed a 27-fold class dataset (containing 3202 sequences), proposed the ACC-Fold method, and obtained an overall accuracy of 87.6% using 5-fold cross-validation. Liu and Hu (2010) constructed a new 27-fold class dataset according to the construction of the Ding and Dubchak dataset (2001). This new dataset contains 1895 sequences with a sequence identity below 35%. Motif frequency, low-frequency power spectral density, amino acid composition, predicted secondary structure, and autocorrelation function values were combined as the set of feature parameters. Using the SVM algorithm and the ensemble classification strategy, the overall accuracy in the independent test was 66.67%. Moreover, studies on datasets consisting of 76, 86, and 199 fold classes have demonstrated improvements (Liu et al., 2012; Dong et al., 2009).

In this study, we reorganized the dataset constructed by Liu et al. (2012). According to the biological characteristics, values of the increment of diversity, motif frequency, predicted secondary structure motifs and the average chemical shift information of predicted secondary structure elements were extracted as feature parameters. Based on the ensemble classification strategy, these combined features were used as the input parameter for the Random Forests algorithm. An independent test and 5-fold cross-validation were used to predict the 76 protein fold classes, which resulted in good protein fold identification. The protein folds of the 27-fold class dataset and the corresponding structural classes were also identified, yielding improved results.

2. Materials and methods

2.1. Protein fold dataset

The 76-fold class dataset constructed by Liu et al. (2012) was reorganized; 8 and 5 protein sequences were added to the training and test set, respectively. Then the training set contains 1744 proteins for training, and the test set contains 1726 proteins for test. The sequence identity of the dataset was below 35%. The number of sequences of each type of protein fold was 10 or greater. The training and test set contained 1744 and 1727 protein chains, respectively. The distribution of the corresponding fold names and sequence numbers is shown in Table 1. The 76-fold class dataset is available at http://202.207.29.245:8080/Ha/HomePage/fzxHomePage.jsp.

Please cite this article in press as: Feng, Z. et al., The recognition of multi-class protein folds by adding average chemical shifts of secondary structure elements. Saudi Journal of Biological Sciences (2015), http://dx.doi.org/10.1016/j.sjbs.2015.10.008

Table 1 Datasets of 76 protein fold classes.

Fold (name)	Ntrain/(Ntest)	Fold (name)	Ntrain/(Ntest)	Fold (name)	Ntrain/(Ntest)
1 (GL)	14/14	27 (ITL)	41/41	53 (SM)	44/44
2 (CY)	10/10	28 (RCD)	13/13	54 (PT-L)	31/31
3 (DB)	92/90	29 (SR)	13/13	55 (PBPI)	26/26
4 (HB)	25/24	30 (F-L)	21/21	56 (CD-L)	7/7
5 (4HC)	8/8	31 (SD)	15/14	57 (L-L)	8/8
6 (EF)	25/23	32 (α-T)	16/16	58 (I-L)	8/7
7 (IL)	86/85	33 (CP)	9/8	59 (C-L)	29/30
8 (CD)	18/18	34 (α-S)	32/33	60 (U-L)	9/8
9 (VCP)	24/24	35 (NRL)	7/7	61 (GRP)	16/16
10 (CLL)	18/17	36 (MC)	9/9	62 (C-DP)	8/9
11 (SH3)	41/41	37 (CFD)	14/14	63 (TED)	26/25
12 (OB)	29/28	38 (C2D)	9/9	64 (DL)	8/9
13 (BT)	11/10	39 (GD)	16/16	65 (ETK)	10/9
14 (TSP)	17/16	40 (PDL)	24/25	66 (BCM)	8/9
15 (LIP)	16/15	41 (AP)	8/8	67 (Z-L)	12/11
16 (TIM)	93/92	42 (PDB)	29/29	68 (S-L)	7/8
17 (FAD)	5/5	43 (6BP)	10/9	69 (ACN)	33/32
18 (FLL)	37/36	44 (7BP)	8/8	70 (PL)	19/19
19 (NAD)	17/16	45 (SR-β)	12/13	71 (Nu)	12/12
20 (P-L)	74/73	46 (DSH)	40/40	72 (Tbp)	18/18
21 (THL)	37/36	47 (β-C)	8/7	73 (DNA)	11/11
22 (RHM)	39/40	48 (AN-α)	13/12	74 (PK)	22/22
23 (HYD)	33/33	49 (HL)	25/26	75 (NH-L)	15/15
24 (PBP)	6/6	50 (RCC)	9/9	76 (CTL)	12/12
25 (β-G)	39/39	51 (P/H)	17/17		
26 (FEL)	101/99	52 (P-L)	12/13		

Note: Ntrain/(Ntest) represents the number of folds in the training/(test) dataset.
Full names: (1) globin-like, (2) cytochrome *c*, (3) DNA-binding 3-helical bundle, (4) 4-helical up-and-down bundle, (5) 4-helical cytokines, (6) EF hand, (7) immunoglobulin-like β-sandwich, (8) cupredoxins, (9) viral coat and capsid proteins, (10) ConA-like lectin/glucanases, (11) SH3-like barrel, (12) OB-fold, (13) β-trefoil, (14) trypsin-like serine proteases, (15) lipocalins, (16) TIM barrel, (17) FAD (also NAD)-binding motif, (18) flavodoxin-like, (19) NAD(P)-binding Rossmann fold, (20) P-loop, (21) thioredoxin-like, (22) ribonuclease H-like motif, (23) hydrolases, (24) periplasmic binding protein-like, (25) β-grasp, (26) ferredoxin-like, (27) small inhibitors/toxins/lectins, (28) RuvA C-terminal domain-like, (29) spectrin repeat-like, (30) ferritin-like, (31) SAM domain-like, (32) α/α toroid, (33) cytochrome P450, (34) α–α superhelix, (35) nuclear receptor ligand-binding domain, (36) multiheme cytochromes, (37) diphtheria toxin/transcription factors/cytochrome *f*, (38) C2 domain-like, (39) galactose-binding domain-like, (40) PDZ domain-like, (41) acid proteases, (42) PH domain-like barrel, (43) 6-bladed β-propeller, (44) 7-bladed β-propeller, (45) single-stranded right-handed β-helix, (46) double-stranded β-helix, (47) β-clip, (48) adenine nucleotide α hydrolase-like, (49) HAD-like, (50) rhodanese/cell cycle control phosphatase, (51) phosphorylase/hydrolase-like, (52) PRTase-like, (53) *S*-adenosyl-L-methionine-dependent methyltransferases, (54) PLP-dependent transferase-like, (55) periplasmic binding protein-like II, (56) cytidine deaminase-like, (57) lysozyme-like, (58) IL8-like, (59) cystatin-like, (60) UBC-like, (61) glyoxalase/bleomycin resistance protein/dihydroxybiphenyl dioxygenase, (62) CBS-domain pair, (63) thioesterase/thiol ester dehydrase-isomerase, (64) dsRBD-like, (65) eukaryotic type KH domain (KH-domain type I), (66) Bacillus chorismate mutase-like, (67) zincin-like, (68) SH2-like, (69) acyl-CoA *N*-acyltransferases (Nat), (70) profilin-like, (71) Nudix, (72) TBP-like, (73) DNA clamp, (74) protein kinase-like (PK-like), (75) Ntn hydrolase-like, and (76) C-type lectin-like.

The first 27 types of the 76 protein fold classes correspond to the dataset of Ding and Dubchak (2001), and each type of fold has been expanded. The number of sequences in the dataset is threefold greater than that of the Ding and Dubchak dataset.

The second dataset used in this study was constructed by Ding and Dubchak. The previously used dataset, with sequence identity below 35%, contained a training set that included 311 protein chains and a test set that included 383 protein chains.

2.2. The selection of feature parameters

2.2.1. Increment of diversity (ID)

The ID algorithm has been successfully used in the classification of protein structure and subcellular localization (Chen and Li, 2007). The ID can be used as a classification prediction algorithm and can extract characteristics of the sequence as parameters of the classification prediction.

In the state space of k dimensions, m_i indicates the absolute frequency of the ith state. The diversity measure for diversity source S:$\{m_1, m_2, \ldots, m_k\}$ is defined as follows:

$$D(K) = M \log M - \sum_{i}^{k} m_i \log m_i \quad (1)$$

Here, $M = \sum_{i=1}^{k} m_i$, $\log(0) = 0$ *if* $n_i = 0$

In this state space, the ID between the source of diversity X $(n_1, n_2, \ldots, n_k)$ and $Y(m_1, m_2, \ldots, m_k)$ is defined as follows:

$$\mathrm{ID}(X, Y) = D(X + Y) - D(X) - D(Y) \quad (2)$$

where $D(X + Y)$, which is termed the combination diversity source space, is the measure of the diversity of the sum of two diversity sources.

Please cite this article in press as: Feng, Z. et al., The recognition of multi-class protein folds by adding average chemical shifts of secondary structure elements. Saudi Journal of Biological Sciences (2015), http://dx.doi.org/10.1016/j.sjbs.2015.10.008

The ID is used to measure the similarity level between two diversity sources. If X is similar to Y, then the value of ID(X, Y) will be small, particularly if $X = Y$, then ID(X, Y) = 0.

Considering the local conservation of fold sequences, the sequence of each protein fold was divided into n segments, and in each segment, the occurrence frequencies of 20 amino acid residues in the protein sequences were extracted as a parameter, as previously described (Chen and Li, 2007; Wang et al., 2014). Thus, the initial parameter of each sequence was converted into a 20*n-dimensional vector that was inputted into the ID algorithm for classification, and an improved result was obtained. Following substantial iterative calculations, when an enzyme sequence was divided into 10 segments, a relatively better result was obtained. Therefore, we selected a 200-dimensional vector as the initial parameter for input into the ID algorithm and obtained 76 ID values for each sequence.

2.2.2. Average chemical shift (ACS)

Several studies have noted that the ACS of a particular nucleus in the protein backbone correlates well to its secondary structure (Sibley and Cosman, 2003; Zhao et al., 2010). Mielke and Krishnan (2003), Mielke and Krishnan (2004), Mielke and Krishnan (2009) have presented a CS-based empirical approach to predict secondary structure and the protein structural class. Arai et al. (2010) have predicted the protein structural class using ^{1}H–^{15}N HSQC spectra. Moreover, CS information has been used to improve the prediction quality for various protein subcellular localizations (Fan and Li, 2012a; Fan and Li, 2012b).

These results suggest that CS information can be regarded as important parameters in the prediction of protein folds. Chemical shift values corresponding to the protein backbone atoms were obtained from the BMRB (http://www.bmrb.wisc.edu) (Seavey et al., 1991). The online web server PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) was used to obtain the predicted secondary structure of each protein sequence in the 76-protein fold class dataset.

We calculated the ACS using a previously described method (Mielke and Krishnan, 2003; Fan and Li, 2012a; Fan and Li, 2012b; Fan et al., 2013; Fan and Li, 2013; Anaika et al., 2003). We selected chemical shift values of $^{1}H_{\alpha}$ and $^{1}H_{N}$ (two types of protein backbone atoms for every amino acid residue of protein sequence P) to calculate the corresponding ACS. Subsequently, each amino acid in the sequence was replaced by its ACS. Following iterative calculations, we selected the averaged chemical shifts of $^{1}H_{\alpha}$ and $^{1}H_{N}$, which were more suitable for predicting protein folds. Protein sequence P is expressed as follows:

$$P = [C_1^i, C_1^i, \cdots C_L^i](i = {}^{1}H_{\alpha}, {}^{1}H_{N}) \tag{3}$$

The auto cross covariance (ACC) (Wold et al., 1993) has been successfully adopted for the prediction of protein folds (Dong et al., 2009; Qi et al., 2015), G-proteins (Guo et al., 2006; Wen et al., 2007), protein interactions (Guo et al., 2008), and β-hairpins (Jun et al., 2010). However, the ACC has primarily been used to study interactions between residues or bases. We are the first to use the ACC at the level of predicted secondary structure elements (helix, strand, or coil) for protein fold prediction (Xinghui et al., 2015). The ACC contains two types of variables: the AC variable measures the correlation between identical properties (i.e., an identical secondary structure element) and the CC variable measures the correlation between different properties. Given the corresponding predicted secondary structure elements (helix, strand, or coil) in one sequence, AC variables describe the average interactions between identical predicted secondary structure elements, and the separation distance between two predicted secondary structure elements is given by lg elements. For example, if two secondary structure elements are neighboring, then $lg = 1$; if the two secondary structure elements are next-to-neighboring, then $lg = 2$, etc. The AC variables were redefined and calculated according to Eq. (4), as follows:

$$\mathrm{AC}\ (i, lg) = \begin{cases} \sum_{j=1}^{L-lg} (S_{i,j} - \bar{S}_i)(S_{i,j+lg} - \bar{S}_i)/(L - lg) & (lg < L) \\ 0 & (lg \geqslant L) \end{cases} \tag{4}$$

Here $\bar{S}_i = \sum_{j=1}^{L} S_{i,j}/L \quad (i = 1, 2, 3)$,where i represents a secondary structure element (helix, strand, or coil), L is the number of secondary structure elements in the protein sequence, and S_{ij} is a feature value of secondary structure element i at position j. $\bar{S}_i$ is the average value for the secondary structure element i along the entire sequence (Zhang et al., 2014).

Given the ACS values for 20 amino acid residues in a sequence, the secondary structure element i contains m residues, and S_{ij} represents the summation of ACS values for m residues.

CC variables were redefined and calculated according to Eq. (5), as follows:

$$\mathrm{CC}\,(i1, i2, lg) = \begin{cases} \sum_{j=1}^{L-lg} (S_{i1,j} - \bar{S}_{i1})(S_{i2,j+lg} - \bar{S}_{i2})/(L - lg) & (lg < L) \\ 0 & (lg \geqslant L) \end{cases} \tag{5}$$

where $i1$ and $i2$ are two different types of secondary structure elements (helix, strand, or coil), and $S_{i1,j}$ is a feature value of secondary structure element $i1$ at position j. $\bar{S}_{i1}$ ($\bar{S}_{i2}$) is the average value for secondary structure element $i1$($i2$) along the entire sequence (Li et al., 2015). The dimension of CC variables is 3*2*lg. The ACC is the summation of variables AC and CC. Following substantial calculations and a comparison of the prediction results, the optimal maximal value of lg was selected as 8 in this study (Zhiwei et al., 2015).

2.2.3. Motif information (M)

A motif is the local conserved region in a protein during evolution (Ben-Hur and Brutlag, 2003) that is often related to biological function. For example, some motifs are related to DNA binding sites and enzyme catalytic sites (Wang et al., 2003). As feature parameters, motif information has been successfully applied for the prediction of superfamilies, protein folds, etc. (Ben-Hur and Brutlag, 2003; Liu et al., 2012; Wang et al., 2014).

Two types of motifs were used in this study: motifs with a biological function obtained by searching the existing functional motif database PROSITE (de Castro et al., 2009) and statistical motifs that were obtained using MEME (http://meme.nbcr.net/meme/cgi-bin/meme.cgi). Motif information (M) includes functional and statistical motifs.

Please cite this article in press as: Feng, Z. et al., The recognition of multi-class protein folds by adding average chemical shifts of secondary structure elements. Saudi Journal of Biological Sciences (2015), http://dx.doi.org/10.1016/j.sjbs.2015.10.008

(1) Functional motif

The PROSITE database was used to obtain protein sequence patterns with notable biological functions. PS_SCAN packets provided by the PROSITE database were used and compiled using a Perl program as a motif-scan tool to search the sequences of the 76-fold class training set, and 181 functional motifs were selected. For an arbitrary sequence in the dataset, the frequencies of different motifs in the sequence were recorded. If a motif occurs once, the corresponding frequency value was recorded as "1"; if the motif occurs twice, the value was recorded as "2", etc.; otherwise if the motif is absent, the corresponding frequency value was recorded as "0". Thus, the frequencies of different functional motifs in a protein sequence were converted into a 181-dimensional vector.

(1) Statistical motif

For statistical motifs, MEME was applied as the motif-scan tool (Bailey et al., 2006). The motifs with the three highest frequencies were selected. Each motif contained 6–10 amino acid residues; thus, 228 motifs were obtained and selected from the 76-fold class training set. For an arbitrary sequence in the dataset, if a motif occurs once, the frequency value was recorded as "1"; if the motif occurs twice, the value was recorded as "2", etc.; otherwise if the motif is absent, the corresponding frequency value was recorded as "0". Thus, frequencies of different statistical motifs in a protein sequence were converted into a 228-dimensional vector.

2.2.4. Predicted secondary structure motifs (P)

Because the protein fold is a description based on the secondary structure, the formation of secondary structure from the sequence influences the folding of the protein. We extracted the occurrence frequencies of three types of predicted secondary structure motifs (P1) from previous studies (Shen and Chou, 2006; Chen and Kurgan, 2007; Yang et al., 2011) as feature parameters, resulting in a 3-dimensional vector. The occurrence frequencies of four types of supersecondary motifs (P2) were subsequently extracted as feature parameters, resulting in a 4-dimensional vector. Finally, the occurrence frequencies of complex supersecondary motifs (P3) were extracted as parameters (Pi represents the three feature sets, with $i = 1$, 2, or 3). Thus, the frequencies of secondary structure motifs, supersecondary motifs, and complex supersecondary motifs were converted into a 15-dimensional vector represented by P. The online web server PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) was used to obtain the predicted secondary structure of each protein sequence. The three feature sets are provided in Table 2.

Table 2 Summary of predicted secondary structure motifs.

Feature set	Occurrence frequencies of the selected features
P1	"E", "C" and "H"
P2	"ECE", "ECH", "HCH" and "HCE"
P3	"ECECE", "ECECH", "ECHCE", "ECHCH" "HCECE", "HCECH", "HCHCE" and "HCHCH"

Note: "H" indicates "helix", "E" indicates "strand", and "C" indicates "coil".

2.3. Random Forests

Random Forests is a classification algorithm developed by Leo Breiman (2001). The general idea of the algorithm is that multiple weak classifiers constitute a strong individual classifier. Random Forests uses a collection of multiple decision trees, in which each decision tree and each split of the decision tree is a classifier, and the final predictions are made by the majority vote of the trees. The advantages of Random Forests include (1) a few parameters to adjust and (2) the data do not require preprocessing. Random Forests uses two important parameters: (1) the number of feature parameters selected by each node of a single decision tree at each split, which is represented by m ($m = \sqrt{M}$, where M is the total number of features that were initially selected), and (2) the number of decision trees, which is represented by k (in this study, $k = 1000$).

The Random Forests algorithm has been successfully used in the prediction of antifreeze proteins (Kandaswamy et al., 2011), DNA-binding residues (Wang et al., 2009), the metabolic syndrome status and β-hairpins (Jia and Hu, 2011). The Random Forests algorithm was applied using R-2.15.1 software (http://www.r-project.org/) and the Random Forest program package.

3. Results and discussion

3.1. Comparison using different parameters

For the 76-fold class dataset, ID, M, P, and ACS values were extracted as feature parameters, with the combined feature vector as input parameters for the Random Forest algorithm. The overall accuracy of the test set in the dataset was 66.69% using an independent test (Fig. 1). As some features and their combinations may give rise to higher accuracies, and in order to know the basis for them to give high accuracies, we also test the effectiveness of the individual features and their various systematic combinations, and the detailed fold-discriminatory accuracies. We then combined the test set with the training set, as previously described (Lin et al., 2013; Shamim et al., 2007; Ghanty and Pal, 2009), and the overall accuracy was 73.43% using 5-fold cross-validation. The identification results from the gradual addition of relevant feature parameters are summarized in Fig. 1.

When only the ID values, which can reflect the local conservation of fold sequences, were used as the feature parameter in the independent test, the overall accuracy was 26.59%. Following the addition of the ACSs of secondary structure elements, the overall accuracy increased to 57.01% (a 30.42% higher overall accuracy). The accuracies for folds 2, 4, 6, etc., increased more than 50%, and the accuracies of folds 1, 3, 11, etc., increased approximately 30%. The accuracies of the remaining folds also improved to varying extents. Note that the ACSs of secondary structure elements substantially affected the identification of protein folds. Furthermore, we can see that the ACSs of secondary structure elements were shown to provide better accuracies than the other individual

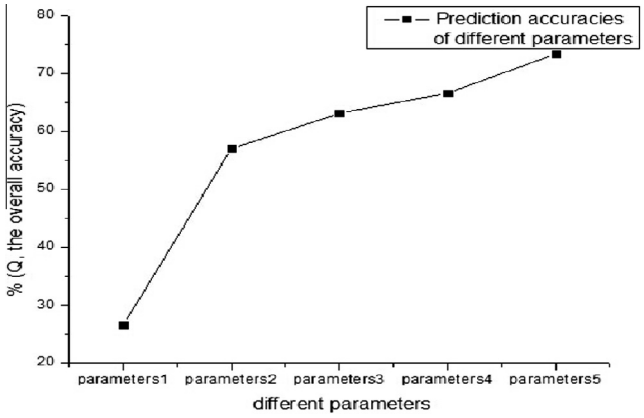

Figure 1 Prediction accuracies for 76 protein fold classes using combinations of different parameters in the test set (%). *Note:* parameter1: ID, increment of diversity values (76 dimensions); parameter2: ID + ACS, values of the increment of diversity and average chemical shifts of secondary structure elements (220 dimensions); parameter3: ID + ACS + M, values of the increment of diversity, average chemical shifts of secondary structure elements and motif frequency (629 dimensions); parameter4: ID + ACS + M + P, values of the increment of diversity, average chemical shifts of secondary structure elements, motif frequency and predicted secondary structure information (644 dimensions); parameter5: ID + ACS + M + P (5-fold cross-validation), values of the increment of diversity, average chemical shifts of secondary structure elements, motif frequency and predicted secondary structure information (644 dimensions); and Q, the overall accuracy.

features. With the specific biological background of protein folds, the proposed feature parameter of ACSs of secondary structure elements was very suitable for predicting 76-fold classes.

Upon the addition of motif frequency information to the values of the ID and ACSs of secondary structure elements, the overall accuracy increased to 63.19%, which represents a 6.18% higher overall accuracy. During this process, the accuracies of folds 2, 10, 14, 40, 49, 50, 60, 71 substantially increased. Furthermore, it was shown that the individual feature of motif frequency information, which reflects the function and structure information of folds, performed very well on the accuracies of folds above. Through investigation on the folds above, the local conservation of the sequences is better than other fold classes, and the sensitivity to motif frequency information is higher.

Finally, addition of the predicted secondary structure motifs, which influence the spatial folding of the protein, resulted in an overall accuracy of 66.69%, and the prediction accuracies of various folds were further improved, resulting in the best overall accuracy (Fig. 1). However, as can be seen, upon the combinations of ACSs of secondary structure elements, motif frequency information and the predicted secondary structure motifs, the overall accuracy was 66.74%, which represents only a 0.05% higher overall accuracy. Overall, as relevant feature parameters were gradually added, the accuracies of a majority of the folds improved to varying extents. The great majority of combinations of features are shown to provide better accuracies than the individual feature. Thus, the combined feature parameters were effective in predicting the 76-fold classes.

For an additional comparison, we combined the training and test set as previously described (Lin et al., 2013; Shamim et al., 2007; Ghanty and Pal, 2009), and the corresponding prediction results using 5-fold cross-validation are summarized in Fig. 1. As can be seen, the overall prediction accuracy using 5-fold cross-validation reached 73.43%, which represents a 6.74% higher overall accuracy. In addition to the 76-protein fold class dataset, the previous results of Liu et al. (2012) using an independent test are also summarized for comparison. Note that the overall accuracy using an independent test was 21.77% higher than that of Liu et al. (2012).

Overall, the results of the 76-protein fold class prediction are encouraging. However, the prediction results for 17, 48, 57, 66 and 67 folds were poor, indicating that future studies are necessary. The web server for protein fold prediction is accessible to the public (http://202.207.29.245:8080/Ha/HomePage/fzxHomePage.jsp).

3.2. Comparison with predictions using the 27-fold class dataset

To evaluate the efficiency of our method, using identical feature parameters, classification strategy, and algorithm, the

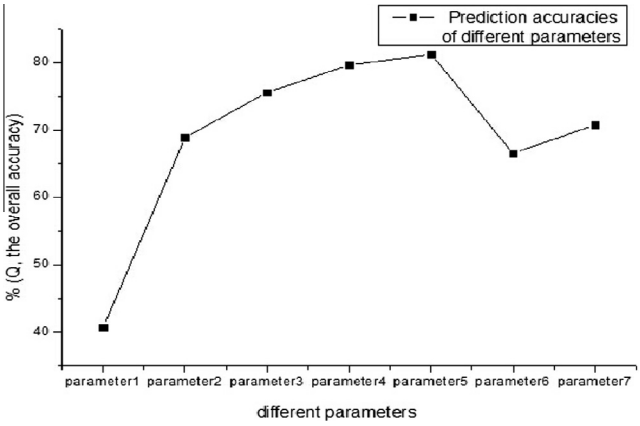

Figure 2 Prediction accuracies of 27 protein fold classes using combinations of different parameters. *Note:* parameter1: ID, increment of diversity values (76 dimensions); parameter2: ID + ACS, values of the increment of diversity and average chemical shifts of secondary structure elements (220 dimensions); parameter3: ID + ACS + M, values of the increment of diversity, average chemical shifts of secondary structure elements and motif frequency (629 dimensions); parameter4: ID + ACS + M + P, values of the increment of diversity, average chemical shifts of secondary structure elements, motif frequency and predicted secondary structure information (644 dimensions); parameter5: ID + ACS + M + P (5-fold cross-validation), values of the increment of diversity, average chemical shifts of secondary structure elements, motif frequency and predicted secondary structure information (644 dimensions); Q, the overall accuracy. The parameter6 summarizes the results of Liu et al. (2012) using an identical dataset. The parameter7 summarizes our results using the dataset constructed by Ding and Dubchak (2001).

Table 3 Identification accuracy using the 27-protein fold class dataset constructed by Ding and Dubchak (%).

Author	Classifier	Accuracy
Ding and Dubchak (2001)	SVM (all-versus-all)	56.0
Chinnasamy et al. (2005)	Tree-augmented naive Bayesian classifier	58.2
Shen and Chou (2006)	OET-KNN	62.1
Nanni (2006)	Fusion of classifiers	61.1
Chen and Kurgan (2007)	PFRES	68.4
Guo and Gao (2008)	GAOEC	64.7
Damoulas and Girolami (2008)	Multi-class multi-kernel	70.0
Zhang et al. (2009)	Increment of diversity	61.1
Ghanty and Pal (2009)	Fusion of different classifiers	68.6
Dong et al. (2009)	ACCFold	70.1
Shen and Chou (2009)	PFP-FunDSeqE	70.5
Yang and Kecman (2011)	MarFold	71.7
Liu et al. (2012)	SVM	69.8
Present study	Random Forests	70.8

first 27-fold classes in the 76-fold class dataset and the dataset constructed by Ding and Dubchak (2001) were also evaluated. Overall accuracies of 79.66% and 70.76%, respectively, for the two datasets were achieved using an independent test (Fig. 2). Moreover, we combined the training and test set of the first 27-fold classes in the 76-fold class dataset and achieved an overall accuracy of 81.21% (which is higher than that of the independent test) using 5-fold cross-validation. The identification results from the gradual addition of relevant feature parameters are summarized in Fig. 2. We also test the effectiveness of the individual features and their various systematic combinations, and the detailed fold-discriminatory accuracies.

Using the identical dataset and test method, the overall accuracy was 13% higher than that of Liu et al. (2012) (Fig. 2), and the prediction using 5-fold cross-validation was superior.

The previous results for the Ding and Dubchak dataset are also summarized in Table 3 for comparison. The accuracy was slightly lower than the best results of Yang et al. (2011), but the overall accuracy in our analysis was higher than previously achieved accuracies (Table 3).

3.3. Identification of the structural classes for the 27-fold classes

As previously described by Shen and Chou (2006), the 27 protein fold classes belong to four structural classes. To evaluate the efficiency of our method, we extracted values of the ID, motif frequency, predicted secondary structure motifs and ACSs of secondary structure elements as feature parameters. The combined feature parameters were used as input parameters for the Random Forests algorithm, and the overall accuracy of the test set for the four structural classes was 93.40% using an independent test. This overall accuracy was 4% higher than the method of Liu et al. (2010) (Table 4). Using this approach, we also evaluated the Ding and Dubchak dataset, which has been used in several studies, and the results were superior to previous results obtained from this dataset (Table 4).

4. Conclusion

Using an identical dataset with different feature parameters can correctly or falsely classify a given protein sequence. Our approach resulted in good predictions and is valid for the following reasons. First, considering the correlation between the biological function of protein folds and secondary structure elements, the composition and combined features of secondary structure elements were adopted as prediction parameters. We additionally calculated the ACSs of secondary structure elements because chemical shifts reflect structural information, such as the nature of hydrogen exchange dynamics, ionization and oxidation states, the influence of the ring current of aromatic residues, hydrogen bonding interactions and long-range correlation information of the sequence. Second, each sequence was divided into segments according to the local conservation of folds, selecting the composition of amino acids as an initial parameter, after which the ID algorithm was further used to obtain ID values as a prediction parameter. Third, motif information, including functional and statistical motifs, was extracted considering the local conservation of kernel structure in the protein folds. Finally, the Random Forests algorithm, as a convenient and highly efficient combination classifier, was employed to yield final classification results that are decided by votes from decision trees.

Table 4 Overall accuracies of structural class identification using different approaches in the test set (%).

Dataset	Author	Structural class				Accuracy
		α	β	α/β	α + β	
Liu et al. (2012)	Present study	95.2	92.91	97.63	84.36	93.40
	Liu and Hu (2010)	97.04	85.43	94.07	78.21	89.24
Ding and Dubchak (2001)	Present study	85.25	88.03	83.22	69.35	82.77
	Liu and Hu (2010)	86.89	88.03	83.22	59.68	81.46
	Zhang et al. (2009)					79.11
	Chinnasamy et al. (2005)					80.52

Please cite this article in press as: Feng, Z. et al., The recognition of multi-class protein folds by adding average chemical shifts of secondary structure elements. Saudi Journal of Biological Sciences (2015), http://dx.doi.org/10.1016/j.sjbs.2015.10.008

Acknowledgements

This work was supported by National Natural Science Foundation of China (30960090, 31260203), The "CHUNHUI" Plan of Ministry of Education, and Talent Development Foundation of Inner Mongolia.

References

Anaika, B.S., Cosman, M., Krishnan, V.V., 2003. An empirical correlation between secondary structure content and averaged chemical shifts in proteins. Biophys. J. 84, 1223–1227.

Arai, H., Tochio, N., Kato, T., Kigawa, T., 2010. An accurate prediction method for protein structural class from signal patterns of NMR spectra in the absence of chemical shift assignments. In: 10th International Conference on Bioinformatics and Bioengineering, 32–37.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S., 2006. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 34, 369–373.

Ben-Hur, A., Brutlag, D., 2003. Remote homology detection, a motif based approach. Bioinformatics 19, 26–33.

Breiman, L., 2001. Random forests. Machine Learn. 45, 5–32.

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

Chen, Y.L., Li, Q.Z., 2007. Prediction of the subcellular location of apoptosis proteins. J. Theor. Biol. 245, 775–783.

Chinnasamy, A., Sung, W., Mittal, A., 2005. Protein structure and fold prediction using tree-augmented naive Bayesian classifier. J. Bioinform. Comput. Biol. 3, 803–819.

Christopher, A.R., Michelle, A.P., 2004. Protein aggregation and neurodegenerative disease. Nat. Med. 10, 10–17.

Damoulas, T., Girolami, M.A., 2008. Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. Bioinformatics 24, 1264–1270.

de Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., Hulo, N., 2009. ScanProsite: detection of ITE signature matches and ProRule associated functional and structural residues in proteins. Nucleic Acids Res. 37, 202–208.

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

Dong, Q.W., Zhou, S.G., Guan, J.H., 2009. A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. Bioinformatics 25, 2655–2662.

Fan, G.L., Li, Q.Z., 2012a. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. J. Theor. Biol. 304, 88–95.

Fan, G.L., Li, Q.Z., 2012b. Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. Amino Acids 43, 545–555.

Fan, G.L., Li, Q.Z., Zuo, Y.C., 2013. Predicting acidic and alkaline enzymes by incorporating the average chemical shift and gene ontology informations into the general form of Chou's PseAAC. Process Biochem. 48, 1048–1053.

Fan, G.L., Li, Q.Z., 2013. Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition. J. Theor. Biol. 334, 45–51.

Ghanty, P., Pal, N.R., 2009. Prediction of protein folds: extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers. IEEE Trans. Nanobiosci. 8, 100–110.

Guo, X., Gao, X., 2008. A novel hierarchical ensemble classifier for protein fold recognition. Protein Eng. Des. Sel. 21, 659–664.

Guo, Y.Z., Yu, L.Z., Wen, Z.N., Li, M.L., 2008. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res. 36, 3025–3030.

Guo, Y., Li, M., Lu, M., Wen, Z., Huang, Z., 2006. Predicting G-protein coupled receptors-G-protein coupling specificity based on autocross-covariance transform. Proteins 65, 55–60.

Jia, S.C., Hu, X.Z., 2011. Using random forest algorithm to predict β-hairpin motif. Protein Pept. Lett. 18, 609–617.

Jun, F.X., Wu, M., You, Z.H., Zhao, X.M., Li, X.L., 2010. Prediction of β-hairpins in proteins using physicochemical properties and structure information. Protein Pept. Lett. 17, 1123–1128.

Kandaswamy, K.K., Chou, K.C., Martinetz, T., Möller, S., Suganthan, P.N., Sridharan, S., Pugalenthi, G., 2011. AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J. Theor. Biol. 270, 56–62.

Krishna, S.S., Grishin, N.V., 2005. Structural drift: a possible path to protein fold change. Bioinformatics 21, 1308–1310.

Li, Z.F., Shi, Y.Y., Ren, H.L., Li, H., Ashraf, M.A., 2015. Simulation of irregular waves in a numerical wave tank. Pol. Maritime Res. S1 22 (86), 21–25.

Li, J.F., Wu, J.G., Chen, K., 2013. PFP-RFSM: protein fold prediction by using random forests and sequence motifs. J. Biomed. Sci. Eng. 6, 1161–1170.

Lindquist, S., Krobitsch, S., Li, L., Sondheimer, N., 2001. Investigating protein conformation-based inheritance and disease in yeast. Philos. Trans. R. Soc. Lond. 356, 169–176.

Lin, C., Zou, Y., Qin, J., Liu, X.R., Jiang, Y., Ke, C.H., Zou, Q., 2013. Hierarchical classification of protein folds using a novel ensemble classifier. PLoS ONE 8, e56499.

Liu, L., Hu, X.Z., 2010. Based on adding combined vectors of motif information and power spectral density for predicting 27-class protein folds. Acta Biophys. Sin. 6, 225–233 (Chinese).

Liu, L., Hu, X.Z., Liu, X.X., Wang, Y., Li, S.B., 2012. Predicting protein fold types by the general form of Chou's pseudo amino acid composition, approached from optimal feature extractions. Protein Pept. Lett. 19, 439–449.

Ma, J., Wollmann, R., Lindquist, S., 2002. Neurotoxicity and neurodegeneration when PrP accumulates in the cytosol. Science 298, 1781–1785.

Ma, J., Lindquist, S., 2002. Conversion of PrP to a self-perpetuating PrPSc-like conformation in the cytosol. Science 298, 1785–1788.

Martin, A.C., Orengo, C.A., Hutchinson, E.G., Jones, S., Karmirantzou, M., Laskowski, R.A., Mitchell, J.B., Taroni, C., Thornton, J. M., 1998. Protein folds and functions. Structure 6, 875–884.

Mielke, S.P., Krishnan, V.V., 2003. Protein structural class identification directly from NMR spectra using averaged chemical shifts. Bioinformatics 19, 2054–2064.

Mielke, S.P., Krishnan, V.V., 2004. An evaluation of chemical shift index-based secondary structure determination in proteins, influence of random coil chemical shifts. J. Biomol. NMR 30, 143–153.

Mielke, S.P., Krishnan, V.V., 2009. Characterization of protein secondary structure from NMR chemical shifts. Prog. Nucl. Magn. Reson. Spectrosc. 54, 141–165.

Ming, L., Haiqiang, L., Xin, N., Ashraf, M.A., 2015. Characteristic studies of micron zinc particle hydrolysis in a fixed bed reactor. Pol. Maritime Res. S1 22 (86), 112–120.

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

Nanni, L., Shi, J.Y., Brahnam, S., Lumini, A., 2010. Protein classification using texture descriptors extracted from the protein backbone image. J. Theor. Biol. 264 (3), 1024–1032.

Qi, D., Feng, J., Li, Y., Liu, A., Hu, J., Xu, H., Ashraf, M.A., 2015. Stability control of propeller autonomous underwater vehicle based on combined sections method. Pol. Maritime Res. S1 22 (86), 157–162.

Seavey, B.R., Farr, E.A., Westler, W.M., Markley, J.L., 1991. A relational database for sequence specific protein NMR data. J. Biomol. NMR 1, 217–236.

Scheibel, T., Bloom, J., Lindquist, S.L., 2004. The elongation of yeast prion fibers involves separable steps of association and conversion. Proc. Natl. Acad. Sci. 101, 2287–2292.

Shamim, M.T., Anwaruddin, M., Nagarajaram, H.A., 2007. Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. Bioinformatics 23, 3320–3327.

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

Shen, H.B., Chou, K.C., 2009. Predicting protein fold pattern with functional domain and sequential evolution information. J. Theor. Biol. 256, 441–446.

Sibley, A.B., Cosman, M., 2003. An empirical correlation between secondary structure content and averaged chemical shifts in proteins. Biophysical Journal 84, 1223–1227.

Thomas, P.J., Qu, B., Pedersen, P.L., 1995. Defective protein folding as a basis of human disease. Elsevier Sci. 20, 456–459.

Wang, X.Y., Schroeder, D., Dobbs, D., Honavar, V., 2003. Automated data-driven discovery of motif-based protein function classifiers. Inf. Sci. 155, 1–18.

Wang, L.J., Yang, M.Q., Yang, J.Y., 2009. Prediction of DNA-binding residues from protein sequence information using random forests. BMC Genomics 10 (Suppl. 1), S1.

Wang, Y., Hu, X.Z., Sun, L.X., Feng, Z.X., Song, H.Y., 2014. Predicting enzyme subclasses by using Random Forest with multicharacteristic parameters. Protein Pept. Lett. 21, 275–284.

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

Wold, S., Jonsson, J., Sjörström, M., Sandberg, M., Rännar, S., 1993. DNA and peptide sequences and chemical processes mutlivariately modelled by principal component analysis and partial least-squares projections to latent structures. Anal. Chim. Acta 277, 239–253.

Xinghui, Y., Linan, L., Xiaomeng, Z., Ashraf, M.A., 2015. Image fusion for travel time tomography inversion. Pol. Maritime Res. S1 22 (86), 149–156.

Yang, T., Kecman, V., Cao, L., Zhang, C., Huang, Z.X., 2011. Margin-based ensemble classifier for protein fold recognition. Expert Syst. Appl. 38, 12348–12355.

Zhang, H.G., Hu, X.Z., Li, Q.Z., 2009. The recognition of 27-Class protein folds: approached by increment of diversity based on multi-characteristic parameters. Protein Pept. Lett. 16, 1112–1119.

Zhao, Y.Z., Alipanahi, B., Li, S.C., Li, M., 2010. Protein secondary structure prediction using NMR chemical shift data. J. Bioinform. Comput. Biol. 8, 867–884.

Zhang, J., Zhang, Z., Ashraf, M.A., 2014. A maximizing aggregate deviation method of multiple attribute decision making. Pak. J. Stat. 30 (6), 623–642.

Zhiwei, Z., Jinzhao, W., Hongyan, T., Ashraf, M.A., Hao, Y., 2015. Approximate equivalence based on symbolic computation and numerical calculation for linear algebra transition systems. Pak. J. Stat. 31 (5), 623–642.

Please cite this article in press as: Feng, Z. et al., The recognition of multi-class protein folds by adding average chemical shifts of secondary structure elements. Saudi Journal of Biological Sciences (2015), http://dx.doi.org/10.1016/j.sjbs.2015.10.008