Journal of Biomolecular Structure and Dynamics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tbsd20

The prediction of protein structural class using averaged chemical shifts

Hao Lin [a] , Chen Ding [a] , Qiang Song [a] , Ping Yang [a] , Hui Ding [a] , Ke-Jun Deng [a] & Wei Chen [b]

[a] Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, 610054, China

[b] Department of Physics, Center for Genomics and Computational Biology, College of Sciences, Hebei United University, Tangshan, 063000, China

Available online: 18 Apr 2012

To cite this article: Hao Lin, Chen Ding, Qiang Song, Ping Yang, Hui Ding, Ke-Jun Deng & Wei Chen (2012): The prediction of protein structural class using averaged chemical shifts, Journal of Biomolecular Structure and Dynamics, 29:6, 643-649

To link to this article: http://dx.doi.org/10.1080/07391102.2011.672628

Journal of Biomolecular Structure and Dynamics
Vol. 29, No. 6, 2012, 643–649

Taylor & Francis
Taylor & Francis Group

The prediction of protein structural class using averaged chemical shifts

Hao Lin[a]*, Chen Ding[a], Qiang Song[a], Ping Yang[a], Hui Ding[a], Ke-Jun Deng[a] and Wei Chen[b]*

[a]*Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China;* [b]*Department of Physics, Center for Genomics and Computational Biology, College of Sciences, Hebei United University, Tangshan 063000, China*

(*Received 1 August 2011; final version received 27 December 2011*)

Knowledge of protein structural class can provide important information about its folding patterns. Many approaches have been developed for the prediction of protein structural classes. However, the information used by these approaches is primarily based on amino acid sequences. In this study, a novel method is presented to predict protein structural classes by use of chemical shift (CS) information derived from nuclear magnetic resonance spectra. Firstly, 399 non-homologue (about 15% identity) proteins were constructed to investigate the distribution of averaged CS values of six nuclei (^{13}CO, ^{13}Cα, ^{13}Cβ, ^{1}HN, ^{1}Hα and ^{15}N) in three protein structural classes. Subsequently, support vector machine was proposed to predict three protein structural classes by using averaged CS information of six nuclei. Overall accuracy of jackknife cross-validation achieves 87.0%. Finally, the feature selection technique is applied to exclude redundant information and find out an optimized feature set. Results show that the overall accuracy increased to 88.0% by using the averaged CSs of ^{13}CO, ^{1}Hα and ^{15}N. The proposed approach outperformed other state-of-the-art methods in terms of predictive accuracy in particular for low-similarity protein data. We expect that our proposed approach will be an excellent alternative to traditional methods for protein structural class prediction.

Keywords: protein structural class; averaged chemical shift; support vector machine

Introduction

The function of a protein is closely associated with its three-dimensional structure (Chou & Zhang, 1992; Yang, Peng, & Chen, 2010). Although the details of the three-dimensional structures of proteins are extremely complicated and irregular, their overall folding patterns are surprisingly simple and regular (Chou & Maggiora, 1998; Feng, Cai, & Chou, 2005). Generally, the globular protein domains can be categorized into all-α, all-β and mixed αβ (including α/β and α + β) according to the types and arrangements of their secondary structural elements (Eisenhaber, Frömmel, & Argos, 1996; Levitt & Chothia, 1976; Orengo et al., 1997). All-α proteins are predominantly composed of α-helices. Correspondingly, all-β proteins are predominantly composed of β-strands. The α/β class represents those proteins in which α-helices and β-strands are largely separated with parallel β-strands, while the α + β class represents those proteins in which α-helices and β-strands are largely mixed with antiparallel β-strands. The knowledge of protein structural class can improve the quality of secondary structure prediction, reduce the scope of conformational searches during energy optimization and provide important information about protein function (Cid, Bunster, Canales, & Gazitua, 1992; Cohen & Kuntz, 1987; Zhang & Chou, 1995; Zhang, Ding, & Chou, 2008).

In the past three decades, many efforts were made for the prediction of protein structural class (Anand, Pugalenthi, & Suganthan, 2008; Cao et al., 2006; Chen, Chen, Zou, & Cai, 2008; Chen, Kurgan, & Ruan, 2008; Chen et al., 2009; Chen, Stach, Homaeian, & Kurgan, 2011; Chou, 1999, 2005; Chou & Cai, 2004; Ding, Zhang, & Chou, 2007; Du, Jiang, He, Li, & Chou, 2006; Gu & Chen, 2009; Gupta, Mittal, & Singh, 2008; Jahandideh, Abdolmaleki, Jahandideh, & Asadabadi, 2007a, 2007b; Kedarisetti, Kurgan, & Dick, 2006; Liao, Liao, Lu, & Cao, 2011; Liu, Zheng, & Wang, 2010a, 2010b; Metfessel, Saurugger, Connelly, & Rich, 1993; Niu, Cai, Lu, Li, & Chou, 2006; Zheng, Li, & Wang, 2010; Zhou, 1998; Zhou & Assa-Munt, 2001). According to the suggestion that the structure class of a protein correlates strongly with its primary sequence (Anand et al., 2008; Chou & Zhang, 1992; Klein, 1986), the amino acid compositions (AAC) were firstly selected as inputs in predictors (Cai, Liu, Xu, & Chou, 2002; Cai, Liu, Xu, & Zhou, 2001; Chou, Liu, Maggiora, & Zhang, 1998; Gu, Chen, & Ni, 2008; Klein & Delisi, 1986; Zhang & Chou, 1992; Zhou, Xu, & Zhang, 1992). However, using AAC to represent a protein would completely lose the sequence-order information. In view of this, the dipeptide and tripeptide compositions were proposed to enhance the predictive power of predictors (Costantini & Facchiano, 2009; Lin & Li, 2007; Yu,

*Corresponding authors. Email: hlin@uestc.edu.cn; greatchen@heuu.edu.cn

ISSN 0739-1102 print/ISSN 1538-0254 online

http://dx.doi.org/10.1080/07391102.2011.672628
http://www.tandfonline.com

Sun, Sang, Huang, & Zou, 2007). As it is well known, the physicochemical properties of 20 amino acids are the important factors in protein folding and can influence the fold mode of proteins. Hence, the pseudo amino acid composition (PseAAC) was developed to describe not only the AACs, but also the long-distance interaction of physicochemical properties between two residues (Cai et al., 2002; Chen, Tian, Zou, Cai, & Mo, 2006; Chen, Zhou, Tian, Zou, & Cai, 2006; Li, Zhou, Dai, & Zou, 2009; Luo, Feng, & Liu, 2002; Sahu & Panda, 2010; Shen, Yang, & Chou, 2006; Xiao, Lin, & Chou, 2008; Zhang & Ding, 2007). Although these features have been very successful in the prediction of protein structural class for high-similarity ($\geqslant 40\%$ identity) data-sets, they were not effective any more for the data-sets with low-similarity ($\leqslant 25\%$ identity) (Yang et al., 2010). To overcome this limitation, the predicted secondary structure information was used to improve predictive performances for low-similarity sequences (Kurgan & Chen, 2007; Kurgan, Cios, & Chen, 2008; Kurgan, Zhang, Zhang, Shen, & Ruan, 2008; Liu & Jia, 2010; Mizianty & Kurgan, 2009; Yang et al., 2010; Zhang, Ding, & Wang, 2011). But, in fact, the features used in secondary structure prediction were usually directly derived from the amino acid sequence. If the secondary structure of a protein chain is not correctly predicted, it will lead to the incorrect prediction of the structural class of this protein. These urge us to mine new parameters for improving predictive performance.

Nuclear magnetic resonance (NMR) spectroscopy plays a unique and important role in the investigation of the structures and dynamics of proteins and other macromolecules due to its ability to provide site-specific information about protein motions over a large range of time scales (Berjanskii & Wishart, 2005). In NMR, the chemical shift (CS) describes the dependence of nuclear magnetic energy levels on the electronic environment in a molecule. CSs are well recognized as powerful indicators of the types of protein structures (Baskaran, Brunner, Munte, & Kalbitzer, 2010). Some works have studied the relations between protein local structures and CSs, and have proved that protein structures are strongly associated with CSs (Cavalli, Salvatella, Dobson, & Vendruscolo, 2007; Szilágyi, 1995; Wishart, 2010; Wishart et al., 2008). Berjanskii and Wishart (2005, 2007) have used secondary CSs to predict protein flexibility. Moreau, Valente, and Almeida (2006) have predicted the amount of secondary structure of proteins based on CSs. Mielke and Krishnan (2003, 2004, 2009) have presented a CS-based empirical approach to predict secondary structure and structural class of proteins. Arai, Tochio, Kato, Kigawa, and Yamamura (2010) have predicted the protein structural class based on ^{1}H and ^{15}N-HSQC spectra. Some encouraging results were obtained. These results suggested that the CS information can be regarded as important parameters to predict protein structural classes.

In this study, we would like to introduce a powerful approach based on support vector machine (SVM) to predict protein structural class by using averaged chemical shift (ACS) information. A low-similarity (~15% identity) protein benchmark data-set was constructed for testing the predictive performance of the proposed method. The feature selection technique was used to find optimal feature set. Jackknife cross-validated results show that the overall accuracy achieves 88.0% with an average sensitivity of 87.3%, suggesting that our method is a promising approach.

Materials and methods

Database

The CS values of nuclei $^{13}\mathrm{C}_{\mathrm{O}}$, $^{13}\mathrm{C}_{\alpha}$, $^{13}\mathrm{C}_{\beta}$, $^{1}\mathrm{H}_{\mathrm{N}}$, $^{1}\mathrm{H}_{\alpha}$ and ^{15}N in proteins were extracted from re-referenced protein chemical shift database (RefDB) (Zhang, Neal, & Wishart, 2003). We initially obtained 2162 re-referenced protein CS files. The following steps were performed to construct a reliable and high quality benchmark data-set. (i) Only proteins in RefDB overlapping with the corresponded Protein Data Bank (PDB) file with sequence identity of 100% were considered. (ii) Only proteins with the annotation of structural class in PDB were considered. (iii) The proteins with less than 50 amino acids were excluded because of lacking enough sequence information for comparison. (iv) The proteins with less than 65% of their residues assigned CSs were excluded. (v) Only proteins with six nuclei assigned CSs were considered. The first three steps can guarantee credible structural information of proteins, and the last two steps can guarantee reliable and abundant CS information. Finally, to avoid any homology bias, the PISCES program (Wang & Dunbrack, 2003) was utilized to remove the highly similarity sequences.

After strictly following the aforementioned procedures, we finally obtained a non-redundant benchmark data-set including 124 all-α, 112 all- β and 163 mixed αβ proteins. Among 399 proteins, 99% (395 sequences) proteins have less than 15% sequence identity; and the sequence identity of the remains ranges from 15 to 25%. This reliable and rigorous data-set provides us a chance to investigate the relations between the CSs and protein structural classes. The benchmark data can also ensure impartially and correct evaluations of the performance of various methods. The protein sequences and ACS information can be freely downloaded from http://cobi.uestc.edu.cn/people/hlin/tools/PSCPred/.

Averaged chemical shift

According to the definition proposed by Mielke and Krishnan (2003), we calculated the ACSs of six nuclei using following formula.

$$\mathrm{ACS}(i) = \sum_{m=1}^{M} \mathrm{CS}(i, m)/M \qquad (1)$$

here $i = {}^{13}\mathrm{C}_{\mathrm{O}}$, $^{13}\mathrm{C}_{\alpha}$, $^{13}\mathrm{C}_{\beta}$, $^{1}\mathrm{H}_{\mathrm{N}}$, $^{1}\mathrm{H}_{\alpha}$ or ^{15}N. M denotes the total number of residues with CS values assigned for nucleus species i. CS(i, m) denotes the CS value of the ith nucleus at the mth residue.

Statistical distribution

The Z-test is a rigorous statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution (Sprinthall, 2003). According to the central limit theorem, many test statistics are approximately normally distributed for large samples. Therefore, these statistical tests can be performed as approximate Z-tests if the sample size is not too small (generally > 30 samples).

The difference of means between any two classes of proteins is measured by Z-score which can be defined by

$$Z = [\text{Mean}_i^{\text{ACS}} - \text{Mean}_j^{\text{ACS}}]/\text{STD} \tag{2}$$

where $\text{Mean}_i^{\text{ACS}}$ and $\text{Mean}_j^{\text{ACS}}$ denote the means of ACSs of ith and jth structural class, respectively. STD means standard deviation which can be calculated by:

$$\text{STD} = \sqrt{s_1^2/n_1 + s_2^2/n_2} \tag{3}$$

here s_1 and s_2 denote standard deviations of samples 1 and 2, respectively. n_1 and n_2 are the numbers of samples 1 and 2, respectively.

Doing multiple two-sample tests would result in an increased chance of committing type-I errors. The analysis of variance (ANOVA) (Sprinthall, 2003) can be used for multi-group means analysis and provides a statistical test of whether or not the means of several groups are all equal. The statistical value, called F-value, is the ratio of sample variance between means square between groups (MSB) and mean square within a group (MSW). The F-value will become larger as the MSB becomes increasingly larger than the MSW. In the absence of differences between groups, the F-value will be near 1.

Algorithm

SVM is a popular supervised machine learning technique. The basic idea of SVM is to map the data of samples into a high-dimensional Hilbert space and to seek a separating hyperplane in this space. In this study, we used free software tool box LibSVM (Fan, Chen, & Lin, 2005) to implement SVM. Because radial basis function (RBF) usually outperforms linear function, polynomial function and sigmoid function, we chose the RBF as the kernel function. The regularization parameter C and kernel parameter γ were tuned to optimize the classification performance using grid search with jackknife cross-validation.

We also examined the predictive performance of other algorithms such as RBFNetwork, J48, NaiveBayes, Meta bagging and Random forest. The free software Weka (Bouckaert et al., 2010) was used to implement these algorithms.

Performance evaluation

In statistical prediction, independent data-set test, subsampling test and jackknife test (or called leave-one-out cross-validation) can be used to examine a predictor for its effectiveness in practical application (Chou & Shen, 2007; Chou & Zhang, 1995). The current study uses jackknife test to examine the predictive accuracy. The following three parameters: sensitivity (Sn), specificity (Sp) and overall accuracy (Oa) are used to evaluate the predictive performance of our approach.

$$\text{Sn} = \text{TP}/(\text{TP} + \text{FN}) \tag{4}$$

$$\text{Sp} = \text{TN}/(\text{TN} + \text{FP}) \tag{5}$$

$$\text{Oa} = [\text{TP}(\alpha) + \text{TP}(\beta) + \text{TP}(\alpha\beta)]/N \tag{6}$$

here TP, FN, TN and FP denote true positives, false positives, true negatives and false positives, respectively. N is total number of sequences.

Results and discussion

Statistical distribution of ACS values

Using the benchmark data-set, we analysed the statistical distribution of ACS values of $^{13}C_O$, $^{13}C_\alpha$, $^{13}C_\beta$, ^1H_N, $^1H_\alpha$ or ^{15}N in three protein structural classes (all α, all β and mixed αβ). The chi-square test demonstrated that the sampling distributions of six-nuclei ACSs obey normal distribution. Thus, Z-test and ANOVA can be used in statistical test. Figure S1 shows histograms and normal distribution graphs of the six ACSs distributions in three classes. As it can be seen from Figure S1, the mean $^{13}C_O$ and $^{13}C_\alpha$ ACS values of all α are 177.02 and 57.89 ppm, which are dramatically larger than that of all β (175.21 and 56.32 ppm). On the contrary, the mean ACSs of $^{13}C_\beta$, ^1H_N, $^1H_\alpha$ and ^{15}N of all α are smaller than those of all β. Because mixed αβ class includes both helices and strands, the mean ACSs of six nuclei are somewhere between the corresponding values of all α and the corresponding values of all β.

For investigating whether the distributions of ACSs of the three protein structural classes are independent of one another, the independent group Z-test was designed to compare the mean of ACSs between arbitrary two protein structural classes. And the ANOVA was used for multi-group means analysis. Table 1 records the Z-scores and F-values of six nuclei. These results quantitatively confirm the trends suggested by the mean values of the data-sets. We observed that the Z-scores and F-values of $^{13}C_O$, $^1H_\alpha$ and ^1H_N are larger than those of $^{13}C_\alpha$, ^{15}N and $^{13}C_\beta$, suggesting that the ACS information of $^{13}C_O$, $^1H_\alpha$ and ^1H_N are more suitable than that of $^{13}C_\alpha$, ^{15}N and $^{13}C_\beta$ for distinguishing the three classes of proteins.

Prediction of protein structural class using ACS

Results in Table 1 show that all *p*-values are $< 10^{-3}$, suggesting that the ACS values of six nuclei are capable of predicting three protein structural classes. Therefore, we examined the accuracy of six nuclei by using SVM algorithm. The overall accuracies are 84.0, 81.2, 76.9, 66.3, 61.9 and 51.9% for $^{13}C_O$, $^1H_\alpha$, ^1H_N, $^{13}C_\alpha$, ^{15}N and $^{13}C_\beta$, respectively. We noticed that the *F*-value of $^1H_\alpha$ ACS (575.1) is larger than that of $^{13}C_O$ ACS (565.9), which implies that $^1H_\alpha$ ACS should achieve higher predicted accuracy. However, the predicted successful rate of $^1H_\alpha$ ACS (81.2%) is lower than that of $^{13}C_O$ ACS (84.0%). Thus, we may draw a conclusion that the statistical results can only influence, but not determine predictive results.

In general, the more parameters the predictor uses, the higher accuracy it will achieve. Therefore, by inputting the ACSs of six nuclei simultaneously into SVM, we achieved an overall accuracy of 87.0% with the average sensitivity of 86.6% (Table 2). This result is better than that of single nucleus. However, some works have demonstrated that information redundancy can reduce the predictive successful rate (Anand et al., 2008; Costantini & Facchiano, 2009; Du et al., 2006; Gu et al., 2008; Sahu & Panda, 2010). For the purpose of improving predictive performance and eliminating redundant information, we examined all combinations of six-nuclei ACSs. Total of 63 experiments ($C_6^1 + C_6^2 + C_6^3 + C_6^4 + C_6^5 + C_6^6 = 63$) are performed for searching the optimized feature set and obtaining the maximum accuracy. All results are recorded in Table SI. Results show that the maximum overall accuracy of 88.0% is achieved with the average sensitivity of 87.3% by using the combination of $^{13}C_O$, $^1H_\alpha$ and ^{15}N ACSs. Here, we noticed that the combination of $^{13}C_O$, $^1H_\alpha$ and ^1H_N ACSs achieves the same accuracy (85.5%) by the combination of $^{13}C_O$ and $^1H_\alpha$. That is to say ^1H_N ACSs does not provide any information to the combination of $^{13}C_O$ and $^1H_\alpha$ for the prediction. Except for ^{15}N ACS, other nuclei ACSs even reduce predictive performance of predictor by combining with $^{13}C_O$ and 1Hα. These demonstrate that the predictive accuracy is influenced by redundant information. Besides, although the accuracy of ^1H_N ACS is higher than that of ^{15}N ACS, the combination of $^{13}C_O$, $^1H_\alpha$ and ^{15}N ACSs can obtain better performance than the combination of $^{13}C_O$, $^1H_\alpha$ and ^1H_N ACSs. This demonstrates again that the statistical results can only influence, but not determine predictive performance.

Comparison with other approach

Some approaches have been developed for predicting protein structural classes. However, due to differences in data and experimental protocol, it is difficult to compare our results with other published results. Here, we performed two comparisons: one is to compare the performance of ACSs with other sequence parameters by using SVM, and the other is to compare the performance of SVM with other algorithms using the same ACS information.

Firstly, we evaluated the accuracies of three kinds of primary sequence parameters. Twenty AAC and 400 dipeptide compositions (DC) are traditional parameters which can reflect not only the composition of protein sequence, but also the order between two amino acids, and have been widely used for protein prediction. The PseAAC is a kind of popular parameter which is able to

Table 1. The statistical test using *Z*-test and ANOVA for ACSs of six nuclei.

	Z-scores			
Sign of nuclei	All α vs. mixed αβ	Mixed αβ vs. all β	All α vs. all β	ANOVA *F*-values
$^{13}C_O$	19.9 ($p < 0.001$)	17.1 ($p < 0.001$)	31.5 ($p < 0.001$)	565.9 ($p < 0.001$)
$^{13}C_\alpha$	12.9 ($p < 0.001$)	8.4 ($p < 0.001$)	18.5 ($p < 0.001$)	195.9 ($p < 0.001$)
$^{13}C_\beta$	5.9 ($p < 0.001$)	6.4 ($p < 0.001$)	11.0 ($p < 0.001$)	63.7 ($p < 0.001$)
^1H_N	15.5 ($p < 0.001$)	13.2 ($p < 0.001$)	25.9 ($p < 0.001$)	349.2 ($p < 0.001$)
$^1H_\alpha$	23.2 ($p < 0.001$)	13.6 ($p < 0.001$)	32.9 ($p < 0.001$)	575.1 ($p < 0.001$)
^{15}N	12.8 ($p < 0.001$)	6.4 ($p < 0.001$)	16.7 ($p < 0.001$)	150.3 ($p < 0.001$)

Table 2. The results of different parameters using SVM.

	All α		All β		Mixed αβ		Average	Average	
Parameters	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp	Oa
400 Dipeptides	0.468	0.836	0.339	0.909	0.656	0.470	0.488	0.739	0.509
400 Dipeptides + 20 AAs	0.492	0.844	0.357	0.906	0.669	0.496	0.506	0.748	0.526
PseAAC	0.685	0.865	0.598	0.854	0.656	0.742	0.647	0.820	0.649
ACSs of $^{13}C_O$	0.847	0.971	0.804	0.941	0.859	0.835	0.836	0.915	0.840
ACSs of six nuclei	0.927	0.975	0.786	0.955	0.883	0.864	0.866	0.931	0.870
$^{13}C_O$, $^1H_\alpha$ and ^{15}N ACSs	0.927	0.986	0.777	0.965	0.914	0.856	0.873	0.936	0.880

Table 3. The results of different algorithms using optimized three-nuclei ACSs.

	All α		All β		Mixed αβ		Average	Average	
Parameters	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp	Oa
SVM	0.927	0.986	0.777	0.965	0.914	0.856	0.873	0.936	0.880
RBFnetwork	0.944	0.975	0.768	0.962	0.890	0.860	0.867	0.932	0.872
J48	0.919	0.971	0.813	0.948	0.859	0.869	0.864	0.929	0.865
NaiveBayes	0.935	0.956	0.804	0.955	0.847	0.873	0.862	0.928	0.862
Meta bagging	0.903	0.967	0.795	0.948	0.853	0.852	0.852	0.922	0.850
Random forest	0.919	0.964	0.777	0.916	0.791	0.852	0.827	0.911	0.829

harbour some sort of sequence order or pattern information. Table 2 records the accuracies of DC, DC combined with AAC and PseAAC. Results show that PseAAC achieves an overall accuracy of 64.9% with an average sensitivity of 64.7% which are superior to another two sequence parameters, but dramatically lower than that of ACSs. These results suggest that the ACS is an outstanding parameter for protein structure prediction.

Subsequently, by use of optimized ACSs, we tested the performances of RBFNetwork, J48, NaiveBayes, Meta bagging and Random forest by using Weka software. Table 3 shows that all algorithms achieve >80% accuracies which proves again that ACS is a kind of excellent parameter for the prediction of protein structural classes. Among the six algorithms, SVM yields the best outcomes. Therefore, we proposed using SVM to perform protein structural class prediction.

On the basis of classification by SCOP database (Andreeva et al., 2008), most works have focused on predicting four classes of protein structural classes: all α, all β, α + β and α/β. The latter two classes are different in the aspect of the secondary structure connectivity, which is considered at a lower level describing topology (Orengo et al., 1997). Thus, we studied and predicted three major classes: all α, all β, mixed αβ according to the classification defined by CATH database (Orengo et al., 1997). Another important reason is that the number of α/β class (10 proteins in benchmark data-set) is too few to have statistical significance. In the future work, we shall collect sufficient α/β proteins to investigate the difference of ACSs between α + β and α/β.

Conclusion

In summary, we have described a promising approach that is capable of rapidly and accurately predicting protein structural classes by using CS information. By the analysis of the distributions of six-nuclei ACSs in three protein structural classes, we found that six-nuclei ACSs are all significantly different among three classes. By using ACSs as parameters, a precise model was proposed. This model can be applicable to proteins of any size for which ^{1}H, ^{13}C and ^{15}N CSs are available. We believed the ACS-based approach will provide novel information for investigating the topology of protein structure.

Acknowledgements

The authors are grateful to the anonymous reviewers for their valuable suggestions and comments, which have led to the improvement of this paper. This work was partially supported by the National Natural Science Foundation of China (11047180), the Scientific Research Foundation of Sichuan Province (2009JY0013) and the Fundamental Research Funds for the Central Universities (ZYGX2009J081).

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.org/10.1080/07391102.2011.672628.

References

Anand, A., Pugalenthi, G., & Suganthan, P.N. (2008). Predicting protein structural class by SVM with class-wise optimized features and decision probabilities. *Journal of Theoretical Biology, 253*, 375–380.

Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J.P., Chothia, C., & Murzin, A.G. (2008). Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Research, 36*, D419–D425.

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

Baskaran, K., Brunner, K., Munte, C.E., & Kalbitzer, H.R. (2010). Mapping of protein structural ensembles by chemical shifts. *Journal of Biomolecular NMR, 48*, 71–83.

Berjanskii, M.V., & Wishart, D.S. (2005). A simple method to predict protein flexibility using secondary chemical shifts. *Journal of the American Chemical Society, 127*, 14970–14971.

Berjanskii, M.V., & Wishart, D.S. (2007). The RCI server: Rapid and accurate calculation of protein flexibility using chemical shifts. *Nucleic Acids Research, 35*, W531–W537.

Bouckaert, R.R., Frank, E., Hall, M.A., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2010). WEKA—Experiences with a Java open-source project. *Journal of Machine Learning Research, 11*, 2533–2541.

Cai, Y.D., Liu, X.J., Xu, X.B., & Chou, K.C. (2002). Prediction of protein structural classes by support vector machines. *Computers & Chemistry, 26*, 293–296.

Cai, Y.D., Liu, X.J., Xu, X.B., & Zhou, G.P. (2001). Support vector machines for predicting protein structural class. *BMC Bioinformatics, 2*, 3.

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

Cavalli, A., Salvatella, X., Dobson, C.M., & Vendruscolo, M. (2007). Protein structure determination from NMR chemical shifts. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 9615–9620.

Chen, C., Chen, L.X., Zou, X.Y., & Cai, P.X. (2008). Predicting protein structural class based on multi-features fusion. *Journal of Theoretical Biology, 253*, 388–392.

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.

Chen, L., Lu, L., Feng, K., Li, W., Song, J., Zheng, L., … Cai, Y. (2009). Multiple classifier integration for the prediction of protein structural classes. *Journal of Computational Chemistry, 30*, 2248–2254.

Chen, K., Stach, W., Homaeian, L., & Kurgan, L. (2011). iFC^2: An integrated web-server for improved prediction of protein structural class, fold type, and secondary structure content. *Amino Acids, 40*, 963–973.

Chen, C., Tian, Y.X., Zou, X.Y., Cai, P.X., & Mo, J.Y. (2006). Using pseudo-amino acid composition and support vector machine to predict protein structural class. *Journal of Theoretical Biology, 243*, 444–448.

Chen, C., Zhou, X., Tian, Y., Zou, X., & Cai, P. (2006). Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. *Analytical Biochemistry, 357*, 116–121.

Chou, K.C. (1999). A key driving force in determination of protein structural classes. *Biochemical and Biophysical Research Communications, 264*, 216–224.

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

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., Liu, W.M., Maggiora, G.M., & Zhang, C.T. (1998). Prediction and classification of domain structural classes. *Proteins, 31*, 97–103.

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

Chou, K.C., & Shen, H.B. (2007). Recent progress in protein subcellular location prediction. *Analytical Biochemistry, 370*, 1–16.

Chou, K.C., & Zhang, C.T. (1992). A correlation-coefficient method to predicting protein-structural classes from amino acid compositions. *European Journal of Biochemistry, 207*, 429–433.

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

Cid, H., Bunster, M., Canales, M., & Gazitua, F. (1992). Hydrophobicity and structural classes in proteins. *Protein Engineering, 5*, 373–375.

Cohen, F.E., & Kuntz, I.D. (1987). Prediction of the three-dimensional structure of human growth hormone. *Proteins, 2*, 162–166.

Costantini, S., & Facchiano, A.M. (2009). Prediction of the protein structural class by specific peptide frequencies. *Biochimie, 91*, 226–229.

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.

Du, Q.S., Jiang, Z.Q., He, W.Z., Li, D.P., & Chou, K.C. (2006). Amino Acid Principal Component Analysis (AAPCA) and its applications in protein structural class prediction. *Journal of Biomolecular Structure & Dynamics, 23*, 635–640.

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

Fan, R.E., Chen, P.H., & Lin, C.J. (2005). Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research, 6*, 1889–1918.

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

Gu, F., & Chen, H. (2009). Evaluating long-term relationship of protein sequence by use of D-interval conditional probability and its impact on protein structural class prediction. *Protein and Peptide Letters, 16*, 1267–1276.

Gu, F., Chen, H., & Ni, J. (2008). Protein structural class prediction based on an improved statistical strategy. *BMC Bioinformatics, 9*, S5.

Gupta, R., Mittal, A., & Singh, K. (2008). A time-series-based feature extraction approach for prediction of protein structural class. *EURASIP Journal on Bioinformatics and Systems Biology*, 235451.

Jahandideh, S., Abdolmaleki, P., Jahandideh, M., & Asadabadi, E.B. (2007a). Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. *Biophysical Chemistry, 128*, 87–93.

Jahandideh, S., Abdolmaleki, P., Jahandideh, M., & Hayatshahi, S.H. (2007b). Novel hybrid method for the evaluation of parameters contributing in determination of protein structural classes. *Journal of Theoretical Biology, 244*, 275–281.

Kedarisetti, K.D., Kurgan, L., & Dick, S. (2006). Classifier ensembles for protein structural class prediction with varying homology. *Biochemical and Biophysical Research Communications, 348*, 981–988.

Klein, P. (1986). Prediction of protein structural class by discriminant analysis. *Biochimica et Biophysica Acta, 874*, 205–215.

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

Kurgan, L., & Chen, K. (2007). Prediction of protein structural class for the twilight zone sequences. *Biochemical and Biophysical Research Communications, 357*, 453–460.

Kurgan, L., Cios, K., & Chen, K. (2008). SCPRED: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. *BMC Bioinformatics, 9*, 226.

Kurgan, L.A., Zhang, T., Zhang, H., Shen, S., & Ruan, J. (2008). Secondary structure-based assignment of the protein structural classes. *Amino Acids, 35*, 551–564.

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

Li, Z.C., Zhou, X.B., Dai, Z., & Zou, X.Y. (2009). Prediction of protein structural classes by Chou's pseudo amino acid composition: Approached using continuous wavelet transform and principal component analysis. *Amino Acids, 37*, 415–425.

Liao, B., Liao, B., Lu, X., & Cao, Z. (2011). A novel graphical representation of protein sequences and its application. *Journal of Computational Chemistry, 32*, 2539–2544.

Lin, H., & Li, Q.Z. (2007). Using pseudo amino acid composition to predict protein structural class: Approached by incorporating 400 dipeptide components. *Journal of Computational Chemistry, 28*, 1463–1466.

Liu, T., & Jia, C. (2010). A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. *Journal of Theoretical Biology, 267*, 272–275.

Liu, T., Zheng, X., & Wang, J. (2010a). Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile. *Biochimie, 92*, 1330–1334.

Liu, T., Zheng, X., & Wang, J. (2010b). Prediction of protein structural class using a complexity-based distance measure. *Amino Acids, 38*, 721–728.

Luo, R.Y., Feng, Z.P., & Liu, J.K. (2002). Prediction of protein structural class by amino acid and polypeptide composition. *European Journal of Biochemistry, 269*, 4219–4225.

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

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. *Journal of Biomolecular NMR, 30*, 143–153.

Mielke, S.P., & Krishnan, V.V. (2009). Characterization of protein secondary structure from NMR chemical shifts. *Progress in Nuclear Magnetic Resonance Spectroscopy, 54*, 141–165.

Mizianty, M.J., & Kurgan, L. (2009). Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. *BMC Bioinformatics, 10*, 414.

Moreau, V.H., Valente, A.P., & Almeida, F.C.L. (2006). Prediction of the amount of secondary structure of proteins using unassigned NMR spectra: A tool for target selection in structural proteomics. *Genetics and Molecular Biology, 29*, 762–770.

Niu, B., Cai, Y.D., Lu, W.C., Li, G.Z., & Chou, K.C. (2006). Predicting protein structural class with AdaBoost Learner. *Protein and Peptide Letters, 13*, 489–492.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., & Thornton, J.M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure, 5*, 1093–1108.

Sahu, S.S., & Panda, G. (2010). A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Computational Biology and Chemistry, 34*, 320–327.

Shen, H.B., Yang, J., & Chou, K.C. (2006). Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition. *Journal of Theoretical Biology, 240*, 9–13.

Sprinthall, R.C. (2003). *Basic statistical analysis* (7th ed.). Boston, MA: Pearson Education Group.

Szilágyi, L. (1995). Chemical Shifts in Proteins Come of Age. *Progress in Nuclear Magnetic Resonance Spectroscopy, 27*, 325–443.

Wang, G., & Dunbrack, R.L. (2003). PISCES: A protein sequence culling server. *Bioinformatics, 19*, 1589–1591.

Wishart, D.S. (2010). Interpreting protein chemical shift data. *Progress in Nuclear Magnetic Resonance Spectroscopy, 58*, 62–87.

Wishart, D.S., Arndt, D., Berjanskii, M., Tang, P., Zhou, J., & Lin, G. (2008). CS23D: A web server for rapid protein structure generation using NMR chemical shifts and sequence data. *Nucleic Acids Research, 36*, W496–W502.

Xiao, X., Lin, W.Z., & Chou, K.C. (2008). Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. *Journal of Computational Chemistry, 29*, 2018–2024.

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, 11*, S9.

Yu, T., Sun, Z.B., Sang, J.P., Huang, S.Y., & Zou, X.W. (2007). Structural class tendency of polypeptide: A new conception in predicting protein structural class. *Physica A: Statistical Mechanics and its Applications, 386*, 581–589.

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

Zhang, C.T., & Chou, K.C. (1995). An analysis of protein folding type prediction by seed-propagated sampling and jackknife test. *Journal of Protein Chemistry, 14*, 583–593.

Zhang, T.L., & Ding, Y.S. (2007). Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. *Amino Acids, 33*, 623–629.

Zhang, T.L., Ding, Y.S., & Chou, K.C. (2008). Prediction protein structural classes with pseudo-amino acid composition: Approximate entropy and hydrophobicity pattern. *Journal of Theoretical Biology, 250*, 186–193.

Zhang, S., Ding, S., & Wang, T. (2011). High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure. *Biochimie, 93*, 710–714.

Zhang, H., Neal, S., & Wishart, D.S. (2003). RefDB: A database of uniformly referenced protein chemical shifts. *Journal of Biomolecular NMR, 25*, 173–195.

Zheng, X., Li, C., & Wang, J. (2010). An information-theoretic approach to the prediction of protein structural class. *Journal of Computational Chemistry, 31*, 1201–1206.

Zhou, G.P. (1998). An intriguing controversy over protein structural class prediction. *Journal of Protein Chemistry, 17*, 729–738.

Zhou, G.P., & Assa-Munt, N. (2001). Some insights into protein structural class prediction. *Proteins, 44*, 57–59.

Zhou, G., Xu, X., & Zhang, C.T. (1992). A weighting method for predicting protein structural class from amino acid composition. *European Journal of Biochemistry, 210*, 747–749.