Amino Acids (2010) 38:721–728
DOI 10.1007/s00726-009-0276-1

ORIGINAL ARTICLE

Prediction of protein structural class using a complexity-based distance measure

Taigang Liu · Xiaoqi Zheng · Jun Wang

Received: 25 November 2008 / Accepted: 11 March 2009 / Published online: 28 March 2009

Abstract Knowledge of structural class plays an important role in understanding protein folding patterns. So it is necessary to develop effective and reliable computational methods for prediction of protein structural class. To this end, we present a new method called NN-CDM, a nearest neighbor classifier with a complexity-based distance measure. Instead of extracting features from protein sequences as done previously, distance between each pair of protein sequences is directly evaluated by a complexity measure of symbol sequences. Then the nearest neighbor classifier is adopted as the predictive engine. To verify the performance of this method, jackknife cross-validation tests are performed on several benchmark datasets. Results show that our approach achieves a high prediction accuracy over some classical methods.

Keywords Symbol sequence complexity · Nearest neighbor algorithm · Jackknife cross-validation test · Performance measure

T. Liu
Department of Applied Mathematics,
Dalian University of Technology, 116024 Dalian, China
e-mail: mathltg@yahoo.com.cn

X. Zheng
College of Advanced Science and Technology,
Dalian University of Technology, 116024 Dalian, China

J. Wang (✉)
Scientific Computing Key Laboratory of Shanghai Universities,
Department of Mathematics, Shanghai Normal University,
200234 Shanghai, China
e-mail: jwang@shnu.edu.cn

Introduction

The structural class is an important attribute widely used to characterize the overall folding type of a protein or its domain (Chou 2005). And the prior knowledge of protein structural class may help to provide useful input for numerous applications that include prediction of protein folding rates, prediction of DNA-binding sites, protein fold recognition, secondary structure content prediction, reduction of the conformation search space, and implementation of a heuristic approach to find tertiary structure (Chen et al. 2008b). With the ever increasing sequence data, there is a great need to develop reliable and effective computational methods to predict protein structural class solely from its primary sequence.

During the past three decades, many different algorithms and efforts have been made to address this problem (Cai and Zhou 2000; Cai et al. 2001, 2002; Chen et al. 2006a, b, 2008a; Chou 1995, 1999, 2000; Jahandideh et al. 2007a, b; Klein et al. 1986; Xiao and Ling 2007; Xiao et al. 2008a, b; Xiao and Wang 2008; Zhang et al. 2008; Zhou 1998). Most of them are based on the amino acid composition (AAC), which was found to be closely related to the structural class of a protein (Nakashima et al. 1986). However, some researchers pointed out that using AAC to represent a protein sample would completely loose the sequence order information (Chou 2001). In view of this, various descriptors that partly took into account the sequence order effects were proposed, e.g., the pseudo amino acid composition (PseAAC) (Chou 2001; Lin and Li 2007; Xiao et al. 2006; Zhang and Ding 2007), polypeptide composition (Luo et al. 2002; Sun and Huang 2006), and function domain composition (Chou and Cai 2004). In addition to investigating different protein representation models, different classification algorithms are also used to implement the structural class prediction methods, such as neural network (Cai and Zhou 2000), support vector machines (SVM) (Anand et al. 2008; Cai et al. 2001, 2002; Chen et al. 2006a, b), fuzzy clustering (Shen et al. 2005), rough sets (Cao et al. 2006), and other complex classification models (Cai et al. 2006; Feng et al. 2005; Kedarisetti et al. 2006). For more details, see a recent review by Chou (2005).

All methods mentioned above need first to choose a set of features to represent a protein sample and then use some machine learning or data mining algorithms as the predictive engine. The performance of these methods relies strongly on the sensitivity and selectivity of the corresponding feature vectors. But when protein sequences are represented by AAC, PseAAC or some other feature descriptors, it is unavoidable that many important features associated with the structural class are partly lost (especially the sequence order information). On the other hand, known that the primary sequence of a protein contains virtually all of the information needed to determine its structure, it is theoretically possible to achieve a much higher prediction accuracy when predicting protein structural class solely from the protein sequence.

In this study, we present a method called NN-CDM, a nearest neighbor classifier with a complexity-based distance measure, to address this problem. Our method bypasses the process of feature extraction and only makes use of a conditional complexity measure of symbol sequences. Among known measures of complexity, the Lempel–Ziv (LZ) complexity measure reflects most adequately the repeated patterns occurring in the sequence (Lempel and Ziv 1976; Gusev et al. 1999), and hence is adopted in this study. This kind of complexity measure was first employed to predict protein subcellular location and structural class by Xiao et al. (2005, 2006), who used it as one component of PseAAC vector. In the present study, instead of comparing the corresponding sequence features, a distance measure based on the LZ complexity is adopted. Then the nearest neighbor (1-NN) algorithm is used to predict the structural class of a test protein. The results show that prediction accuracies are significantly improved.

Materials and methods

Structural classes

According to the definition by Levitt and Chothia (1976), a protein of known structure usually can be classified into one of the following four structural classes: all-α, all-β, α/β and $\alpha + \beta$. This classification scheme has been modified repeatedly by changing the thresholds for the amount of helices and strands. It is generally believed that the SCOP classification is more natural and provides more reliable information to study protein structural classes when compared with the other classifications (Chou and Maggiora 1998; Kedarisetti et al. 2006). Currently, besides the above four classes, there are also seven classes listed in the SCOP database , i.e., (1) multi-domain proteins; (2) membrane and cell surface proteins; (3) small proteins; (4) coiled coils proteins; (5) low-resolution proteins; (6) peptides; and (7) designed proteins. In the present study, our research focuses on computational prediction of the major four categories at the first level of hierarchy, as they include the great majority of protein sequences and are the basis for most comparable approaches.

Datasets

Two widely studied protein datasets constructed by Zhou (1998) are used as the working datasets to demonstrate the performance of the proposed algorithm. The first dataset contains 277 domains (Z277) and the second one consists of 498 domains (Z498). In addition, in order to further evaluate the reliability of the new method, the other two datasets constructed by Chou and Maggiora (1998) (CM359) and Chou (1999) (C204), respectively are also studied separately. More details about the four datasets are listed in Table 1.

Distance measure based on the LZ decomposition of symbol sequences

To perform the nearest neighbor algorithm, one should first define a distance measure between given proteins. In this study, a certain complexity measure from information theory is adopted.

Let $\mathcal{A}$ be a finite alphabet. A sequence S of length n on $\mathcal{A}$ is an ordered n-tuple $S = s_1 s_2 \ldots s_n$, where $s_i \in \mathcal{A}$. Let $S[i{:}j]$ be the substring of S that starts at position i and ends at position j, i.e., $S[i{:}j] = s_i s_{i+1} \ldots s_j$ for $1 \le i \le j \le n$. The *LZ complexity* of a nonempty sequence S, denoted by $c(S)$, is defined as the minimal number of steps in a certain (optimal) procedure of its synthesis

$$S = S[1 : i_1]S[i_1 + 1 : i_2] \cdots S[i_{k-1} + 1 : i_k] \cdots S[i_{m-1} + 1 : n],$$

where $S[i_{k-1} + 1{:}i_k]$ is a fragment (component) generated at the kth step. And at each step, only two operations are

Table 1 The compositions of four datasets in this study

Dataset	All-α	All-β	α/β	$\alpha + \beta$	Total
Z277	70	61	81	65	277
Z498	107	126	136	129	498
CM359	82	85	99	93	359
C204	52	61	45	46	204

allowed: either copying the longest fragment from the part of S that has already been synthesized or generating an additional symbol which ensures the uniqueness of each component $S[i_{k-1} + 1{:}i_k]$. Lempel and Ziv called the complexity decomposition of a sequence S based on the rule above the *exhaustive history* of S, denoted by $H(S)$, and proved that every sequence S has a unique exhaustive history (Lempel and Ziv 1976). For example, for the sequence $S =$ AEFFGEFFGAE, its exhaustive history is $H(S)$=A·E·F·FG·EFFGA·E, where "·" is used to separate the decomposition components. So $c(S) = 6$.

For any sequences S and Q, let SQ be the concatenation of S and Q. Here, S is called a prefix of SQ and SQ is called an extension of S. It has been proven that, the inequality $c(SQ) - c(S) \leq c(Q)$ is always valid. This shows that the steps required to extend S to SQ are always less than the steps required to build Q from an empty sequence. Meanwhile, the more similar sequences S and Q are, the smaller $c(SQ) - c(S)$ is (Otu and Sayood 2003). Accordingly, a dissimilarity measure can be defined as

$$d(S,Q) = c(SQ) - c(S).$$

Note that $d(S, Q) \neq d(Q, S)$. To ensure the symmetry condition and eliminate the effect of different sequence lengths, the final distance between S and Q is defined as

$$d^*(S,Q) = \frac{\max\{d(S,Q), d(Q,S)\}}{\max\{c(S), c(Q)\}}.$$

To confirm the validity of d^*, we should examine whether it fulfils the three axioms of a metric. The symmetry condition is certainly satisfied. Known that $c(SQ) \geq c(S)$, the positive axiom of a metric is also ensured. However, according to above definitions, $c(SS) = c(S)$ or $c(S) + 1$, the identity condition of the distance metric does not hold when $c(SS) \neq c(S)$. We here remedy it by redefining $d(S, Q) = 0$ if $S = Q$. The triangle inequality is also satisfied up to an additional small error term (Otu and Sayood 2003).

The nearest neighbor algorithm

Protein structural class prediction is usually formulated as a multi-class classification problem. In this study, we employ the nearest neighbor algorithm as the predictor. The nearest neighbor algorithm is a simple nonparametric classification algorithm. Explicitly, a query sequence is assigned to a prior known category that is found most similar to it in terms of a distance/similarity measure (in the present paper, we use d^*). A natural generalization of the nearest neighbor algorithm is the so-called k-nearest neighbors (k-NN) algorithm, where the k-nearest samples are selected and the query sequence is assigned to the category most frequently represented among them. In detail, given a test protein S of unknown category, this algorithm first finds the k-nearest neighbors in the training set $\{S_i\}$ ($i = 1,2,\ldots,N$), where N is the number of training sequences. Then it assigns a prediction label to the test sample S according to the categories of its neighbors. For example, we have four categories, C_1, C_2, C_3 and C_4. Denote the number of k-NN in each category by N_1, N_2, N_3 and N_4, respectively. If

$$N_1 = \max\{N_1, N_2, N_3, N_4\},$$

the sample S should be classified to the category C_1. If there are two or more maximum numbers, one can compare their relative orders in the k-NN.

Despite its simplicity, the nearest neighbor algorithm can give competitive performance compared with many other methods. It has been widely used in bioinformatics, including prediction of protein secondary structure (Yi and Lander 1993), protein β-turn (Kim 2004), protein subcellular location (Cai and Chou 2003), etc.

Performance measures

The performance of our method is verified through the leave-one-out test, i.e., the jackknife test. During the process of the jackknife test, each protein sequence in the dataset is singled out in turn as a test sample, and the remaining protein sequences are used as a training dataset to predict its structural class. Compared with other cross-validation methods, such as the sub-sampling and self-consistency test, the jackknife test is considered to be the most objective way (Chou and Shen 2007) and hence has been increasingly and widely used to examine the performance of various predictors (Cai and Zhou 2000; Cai et al. 2001, 2002; Jahandideh et al. 2007a, b; Zhang et al. 2008).

To evaluate the performance of our method comprehensively, we report standard performance measures over each structural class, including sensitivity (recall or accuracy), specificity, precision, false positive rate (FPR) and Matthews correlation coefficient (MCC). As reported by King and Guda (2007), the sensitivity for class C_j, denoted by Sens$_j$, is defined as the fraction of proteins belonging to class C_j that are correctly predicted; the specificity for class C_j, denoted by Spec$_j$, is defined as the fraction of proteins not in class C_j that are correctly predicted; the precision for class C_j, denoted by Prec$_j$, is defined as the fraction of proteins predicted to be in class C_j that are correct predictions; the FPR for class C_j, denoted by FPR$_j$, is defined as the fraction of proteins not in class C_j that is incorrectly predicted to be in class C_j. MCC provides a single measure of evaluating specificity and sensitivity together, ranging from -1 to 1, where it equals 1 for perfect predictions and 0 for random assignments, and less than 0 if predictions are worse than random guessings (Matthews 1975). Finally, overall accuracy is defined as the fraction of the proteins tested that are classified correctly. Explicitly, they are defined by the following formulas:

$$\mathrm{Sens}_j = \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FN}_j} = \frac{\mathrm{TP}_j}{|C_j|},$$

$$\mathrm{Spec}_j = \frac{\mathrm{TN}_j}{\mathrm{TN}_j + \mathrm{FP}_j} = \frac{\mathrm{TN}_j}{\sum_{k \neq j} |C_k|},$$

$$\mathrm{Prec}_j = \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FP}_j},$$

$$\mathrm{FPR}_j = \frac{\mathrm{FP}_j}{\mathrm{TN}_j + \mathrm{FP}_j} = \frac{\mathrm{FP}_j}{\sum_{k \neq j} |C_k|},$$

$$\mathrm{MCC}_j = \frac{\mathrm{TP}_j \mathrm{TN}_j - \mathrm{FP}_j \mathrm{FN}_j}{\sqrt{(\mathrm{TP}_j + \mathrm{FP}_j)(\mathrm{TP}_j + \mathrm{FN}_j)(\mathrm{TN}_j + \mathrm{FP}_j)(\mathrm{TN}_j + \mathrm{FN}_j)}},$$

$$\text{Overall accuracy} = \frac{\sum_j \mathrm{TP}_j}{\sum_j |C_j|},$$

where TP_j, TN_j, FP_j, FN_j, and $|C_j|$ are the number of true positives, true negatives, false positives, false negatives, and proteins in the structural class C_j, respectively.

Results and discussion

Evaluation of different size *k*-NN

In this section, we determine the optimal value of *k* by evaluating the prediction performance of our method over different size *k*-NN models (up to 20-NN). Our results show that 1-NN model as well as 2-NN model achieves the highest overall accuracies on the two standard datasets, 85.2% for Z277 and 93.8% for Z498, respectively (Fig. 1). Therefore, the results reported in the rest of this study use the 1-NN model, unless otherwise stated.

Prediction performance using 1-NN model

The results obtained by our method are shown in Table 2. Referring to Table 2, specificity is high across all classes (from 92.3 to 96.4%), whereas sensitivity ranges from 83.6 to 93.8%, with the exception of $\alpha + \beta$ class (69.2%) for the first dataset (Z277). All-α class has the highest FPR (7.70%) and its sensitivity is very high. This result shows that all-α class is often prone to over-prediction. Although $\alpha + \beta$ class has the lowest sensitivity (69.2%), its precision is very good (81.8%), which is typical when a class is under-prediction. Among the four structural classes, α/β class has the highest sensitivity, specificity, precision and MCC. This may be related to the proportion of α/β class in the training set in which α/β class occupies the biggest part, as shown in Table 1. Generally speaking, the more positive samples in the training set, the more accurate the prediction by the nearest neighbor algorithm would theoretically be. For the second dataset (Z498), similar phenomena can also be seen from Table 2. Furthermore, there is an improvement in varying degrees in terms of sensitivity, specificity, precision and MCC across all classes for the second dataset (Z498). This is probably because that this dataset has more protein samples than the first dataset (Z277). If possible, we can obtain more accurately predicted results through constructing as large a dataset as possible to train a *k*-NN classifier.

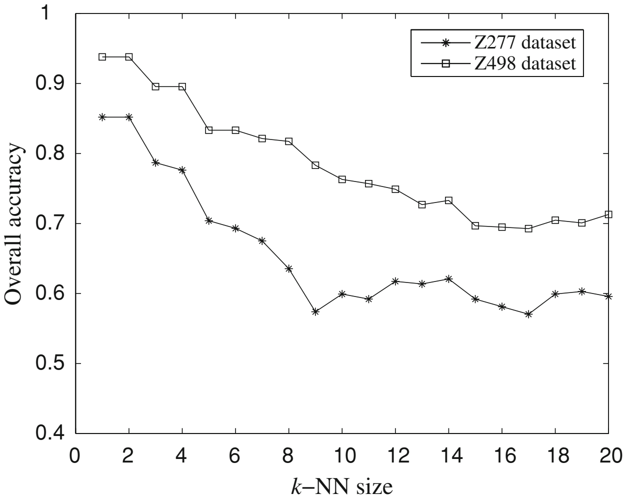

Fig. 1 This graph shows how different values of *k* affect the overall accuracies of our method on the two datasets

Comparison with other methods

To evaluate the performance of our method, we make comparisons with some of existing methods on the same datasets and choose recall and overall accuracy values as the comparative measures. These methods include component coupled algorithm (Zhou 1998), neural network method (Cai and Zhou 2000), SVM (Cai et al. 2001), rough sets (Cao et al. 2006), LogitBoost method (Feng et al. 2005), VPMCD (Raghuraj and Lakshminarayanan 2008), IGA-SVM (Li et al. 2008), SVM fusion network (Chen et al. 2006b), two-stage hybrid neural discriminant model (Jahandideh et al. 2007a). The results by the jackknife test are listed in Table 3.

Table 3 shows that our method achieves the best performance among these methods for the first dataset (Z277), with an overall accuracy of 85.2%. And the overall accuracy of our method is 93.8% for the second dataset (Z498), which is only slightly lower than that of LogitBoost. However, we should point out that the LogitBoost method,

Table 2 Results for the two datasets using 1-NN model

Dataset	Structural class	Sensitivity	Specificity	Precision	FPR	MCC	Overall accuracy
Z277	All-α	91.4	92.3	80.0	7.70	80.3	85.2
	All-β	83.6	96.3	86.4	3.70	80.9	
	α/β	93.8	96.4	91.6	3.60	89.6	
	$\alpha+\beta$	69.2	95.3	81.8	4.7	68.5	
Z498	All-α	96.3	95.9	86.6	4.10	88.9	93.8
	All-β	93.7	98.9	96.7	1.10	93.6	
	α/β	95.6	98.9	97.0	1.10	94.9	
	$\alpha+\beta$	89.9	98.1	94.3	1.90	89.4	

Table 3 Comparison of different methods by the jackknife test for the Z277 and Z498 datasets

Dataset	Method	Recall for each class (%)				Overall accuracy (%)
		All-α	All-β	α/β	$\alpha + \beta$	
Z277	Component coupled (Zhou, 1998)	84.3	82.0	81.5	67.7	79.1
	Neural network (Cai and Zhou, 2000)	68.6	85.2	86.4	56.9	74.7
	SVM (Cai et al., 2001)	74.3	82.0	87.7	72.3	79.4
	Rough sets (Cao et al., 2006)	77.1	77.0	93.8	66.2	79.4
	LogitBoost (Feng et al., 2005)	81.4	88.5	92.6	72.3	84.1
	VPMCD (Raghuraj and Lakshminarayanan, 2008)	85.7	85.0	92.9	84.4	84.2
	IGA-SVM (Li et al., 2008)	84.3	88.5	92.6	70.7	84.5
	Our method	91.4	83.6	93.8	69.2	85.2
Z498	Component coupled (Zhou, 1998)	93.5	88.9	90.4	84.5	89.2
	Neural network (Cai and Zhou, 2000)	86.0	96.0	88.2	86.0	89.2
	SVM (Cai et al., 2001)	88.8	95.2	96.3	91.5	93.2
	Rough sets (Cao et al., 2006)	87.9	91.3	97.1	86.0	90.8
	LogitBoost (Feng et al., 2005)	92.6	96.0	97.1	93.0	94.8
	SVM fusion network (Chen et al., 2006b)	99.1	96.0	80.9	91.5	91.4
	Two-stage hybrid neural discriminant model (Jahandideh et al., 2007a)	95.3	88.9	94.1	93.0	92.8
	Our method	96.3	93.7	95.6	89.9	93.8

which combines many weak classifiers together to build up a strong classifier, is theoretically complicated and time-consuming in the training phase. In contrast to LogitBoost, our method is more straightforward and simpler, i.e., the adopted nearest neighbor algorithm is a nonparametric classification algorithm and has no training-required. Moreover, the pairwise sequence distance is directly evaluated by the LZ complexity measure rather than extracting features from protein sequences.

In order to test current method strictly and investigate the effect of homology on the performance of the new method, the other two datasets are also studied separately. As reported by some researchers, the CM359 dataset has the high sequence similarity. On the contrary, the average sequence similarity in each structural class is lower than 30% (Cai et al. 2002; Lin and Li 2007) in the C204 dataset. The results by the jackknife test are listed in Table 4, compared with several prior works for the same datasets.

It can be seen from Table 4 that for the CM359 dataset, the overall jackknife success rate obtained by our method is 97.1%, which is slightly higher than the results of Kurgan and Homaeian (2006). For the C204 dataset, the overall accuracy of the current method is 91.2%, which is the same with that of the Binary-tree SVM method, and IGA-SVM method proposed by Li et al. (2008) achieves the highest overall predictive accuracy 99.5%. However, it should be pointed out that IGA-SVM coupled the improved genetic algorithm (IGA) with the SVM to predict protein structural classes. This improved GA was applied to the selection of an optimized feature subset and the optimization of SVM parameters (Li et al. 2008). In other words, it would pay very high computational cost for the good prediction

Table 4 Comparison of different methods by the jackknife test for the CM359 and C204 datasets

Dataset	Method	Recall for each class (%)				Overall accuracy (%)
		All-α	All-β	α/β	$\alpha + \beta$	
CM359	Geometric classifier (Chou and Maggiora 1998)	89.0	83.5	85.9	78.5	84.1
	Geometric classifier (Bu et al. 1999)	89.0	78.8	84.9	86.0	84.7
	Component Coupled (Bu et al. 1999)	92.7	90.6	85.7	93.6	90.5
	SVM (Cai et al. 2003)	92.7	96.5	94.9	96.8	95.3
	StackingC ensemble (Kedarisetti et al. 2006)	–	–	–	–	96.4
	Instance-based classifier (Kurgan and Homaeian 2006)	–	–	–	–	97
	SVM (Kurgan and Homaeian 2006)	–	–	–	–	97
	Our method	96.2	97.6	98.9	95.6	97.1
C204	Second-order component-coupled algorithm (Chou 1999)	–	–	–	–	77
	SVM (Cai et al. 2002)	75	90	64	64	74.5
	Supervised fuzzy clustering (Shen et al. 2005)	73.1	90.2	62.2	63.1	73.5
	LogitBoost (Cai et al. 2006)	90.4	88.5	80.0	73.9	83.8
	Augmented covariant discriminant algorithm (Xiao et al. 2006)	82.7	90.2	100	87.0	89.7
	SVM (Chen et al. 2006a)	88.5	96.7	77.8	73.9	85.3
	Binary-tree SVM (Zhang and Ding 2007)	90.4	100	97.8	73.9	91.2
	Multi-features fusion (Chen et al. 2008a)	92.3	93.4	95.6	78.3	90.2
	IGA-SVM (Li et al. 2008)	100	100	97.8	100	99.5
	WSVM (Qiu et al. 2008)	86.5	82.0	91.1	91.3	87.3
	Our method	88.5	100	97.8	76.1	91.2

performance. Compared with IGA-SVM, our method is more straightforward and simpler. The results also indicate that the overall accuracy by our method is about 16 and 6% than those of the two SVMs, which are based on AAC and PseAAC, respectively. Meanwhile, it is worth noting that the Augmented covariant discriminant algorithm, which incorporates the LZ complexity factor of a protein sequence as one component of its PseAAC, also achieves a very high overall accuracy (89.7%). This result reveals that information reflected by the LZ decomposition process contains the essential sequence patterns associated with the structural class. In summary, the above results indicate that our method is very promising and may at least play an important complementary role to the other existing methods.

Conclusion

In this study, we present a new method, NN-CDM, to predict protein structural class solely from sequence information. NN-CDM combines the nearest neighbor algorithm with a complexity-based distance measure. The jackknife cross-validation test is performed on four benchmark datasets. Results show that our method presents a satisfying prediction accuracy compared with other existing methods and may offer a cost-effective alternative to predict protein structural class. In addition, because of the generality of this method, it can be easily extended to the prediction of other protein attributes in the future, such as subcellular location, membrane protein type, enzyme family classification and so on.

The characteristics of our method can be concluded into the following four aspects. First, the nearest neighbor algorithm is adopted as the predictive engine in our method. It is a nonparametric classification algorithm and has no training-required. Second, our method bypasses the process of feature extraction and avoids the bias of selecting features. Third, it indirectly uses specific patterns associated with the structural class by the LZ decomposition of protein sequences. Fourth, along with the third characteristic, sequence-order effects are partly incorporated into the prediction model. However, our method also suffers from the disadvantage of high computational load relative to word statistic-based methods (computational complexity of the LZ decomposition algorithm is $O(n^2)$, where n is the sequence length). But we strongly believe that accuracy is by far much more important than speed in protein structural class prediction because the latter can be easily solved by the rapid improvement in CPU performance. On the other hand, it is expected to develop effective algorithms for pairwise distance measures to help to alleviate this computation limitation.

Acknowledgments This work was supported in part by Leading Academic Discipline Project of Shanghai Normal University (No. DZL803) and Shanghai Leading Academic Discipline Project (No. S30405).

References

Anand A, Pugalenthi G, Suganthan PN (2008) Predicting protein structural class by SVM with class-wise optimized features and decision probabilities. J Theor Biol 253:375–380

Bu WS, Feng ZP, Zhang ZD, Zhang CT (1999) Prediction of protein (domain) structural classes based on amino-acid index. Eur J Biochem 266:1043–1049

Cai YD, Chou KC (2003) Nearest neighbor algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochem Biophys Res Commun 305:407–411

Cai YD, Zhou GP (2000) Prediction of protein structural classes by neural network. Biochimie 82:783–785

Cai YD, Liu XJ, Xu XB, Zhou GP (2001) Support vector machines for predicting protein structural class. BMC Bioinform 2:1–5

Cai YD, Liu XJ, Xu XB, Chou KC (2002) Prediction of protein structural classes by support vector machines. Comput Chem 26:293–296

Cai YD, Liu XJ, Xu XB, Chou KC (2003) Support vector machines for prediction of protein domain structural class. J Theor Biol 221:115–120

Cai YD, Feng KY, Lu WC, Chou KC (2006) Using logitboost classifier to predict protein structural classes. J Theor Biol 238:172–176

Cao YF, Liu S, Zhang L, Qin J, Wang J, Tang KX (2006) Prediction of protein structural class with rough sets. BMC Bioinform 7:1–6

Chen C, Tian YX, Zou XY, Cai PX, Mo JY (2006a) Using pseudoamino acid composition and support vector machine to predict protein structural class. J Theor Biol 243:444–448

Chen C, Zhou XB, Tian YX, Zou XY, Cai PX (2006b) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357:116–121

Chen C, Chen LX, Zou XY, Cai PX (2008a) Predicting protein structural class based on multi-features fusion. J Theor Biol 253:388–392

Chen K, Kurgan LA, Ruan JS (2008b) Prediction of protein structural class using novel evolutionary collocation-based sequence representation. J Comput Chem 29:1596–1604

Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 21:319–344

Chou KC (1999) A key driving force in determination of protein structural classes. Biochem Biophys Res Commun 264:216–224

Chou KC (2000) Prediction of protein structural classes and subcellular locations. Curr Protein Pept Sci 1:171–208

Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43:246–255

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

Chou KC, Cai YD (2004) Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun 321:1007–1009

Chou KC, Maggiora GM (1998) Domain structural class prediction. Protein Eng 11:523–538

Chou KC, Shen HB (2007) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16

Feng KY, Cai YD, Chou KC (2005) Boosting classifier for predicting protein domain structural class. Biochem Biophys Res Commun 334:213–217

Gusev VD, Nemytikova LA, Chuzhanova NA (1999) On the complexity measures of genetic sequences. Bioinformatics 15:994–999

Jahandideh S, Abdolmaleki P, Jahandideh M, Asadabadi EB (2007a) Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophys Chem 128:87–93

Jahandideh S, Abdolmaleki P, Jahandideh M, Hayatshahi SHS (2007b) Novel hybrid method for the evaluation of parameters contributing in determination of protein structural classes. J Theor Biol 244:275–281

Kedarisetti KD, Kurgan L, Dick S (2006) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348:981–988

Kim S (2004) Protein β-turn prediction using nearest-neighbor method. Bioinformatics 20:40–44

King BR, Guda C (2007) ngLOC: an n-gram-based Bayesian method for estimating the subcellular proteomes of eukaryotes. Genome Biol 8:R68

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

Kurgan L, Homaeian L (2006) Prediction of structural classes for protein sequences and domainsImpact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recognit 39:2323–2343

Lempel A, Ziv J (1976) On the complexity of finite sequence. IEEE T Inform Theory 22:75–81

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

Li ZC, Zhou XB, Lin YR, Zou XY (2008) Prediction of protein structure class by coupling improved genetic algorithm and support vector machine. Amino Acids 35:581–590

Lin H, Li QZ (2007) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 28:1463–1466

Luo RY, Feng ZP, Liu JK (2002) Prediction of protein structural class by amino acid and polypeptide composition. Eur J Biochem 269:4219–4225

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451

Nakashima H, Nishikawa K, Ooi T (1986) The folding type of a protein is relevant to the amino acid composition. J Biochem 99:153–162

Otu HH, Sayood K (2003) A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19:2122–2130

Qiu JD, Luo SH, Huang JH, Liang RP (2008) Using support vector machines for prediction of protein structural classes based on discrete wavelet transform. J Comput Chem. doi:10.1002/jcc.21115

Raghuraj R, Lakshminarayanan S (2008) Variable predictive model based classification algorithm for effective separation of protein structural classes. Comput Biol Chem 32:302–306

Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 334:577–581

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30:469–475

Xiao X, Ling WZ (2007) Using cellular automata images to predict protein structural classes. In: The 1st international conference on bioinformatics and biomedical engineering. Wuhan, pp 352–355

Xiao X, Wang P (2008) Predict of protein structural classes based on gray-level co-occurrence matrix feature of protein CAI. In: The

2nd international conference on bioinformatics and biomedical engineering. Shanghai, pp 220–223

Xiao X, Shao SH, Ding YS, Huang ZD, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28:57–61

Xiao X, Shao SH, Huang ZD, Chou KC (2006) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27:478–482

Xiao X, Li WZ, Chou KC (2008a) Using grey dynamic modeling and pseudo amino acid components to predict protein structural classes. J Comput Chem 29:2018–2024

Xiao X, Wang P, Chou KC (2008b) Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. J Theor Biol 254:691–696

Yi TM, Lander ES (1993) Protein secondary structure prediction using nearest-neighbor methods. J Mol Biol 232:1117–1129

Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids 33:623–629

Zhang CT, Chou KC, Maggiora GM (1995) Predicting protein structural classes from amino acid composition: application of fuzzy clustering. Protein Eng 8:425–435

Zhang TL, Ding YS, Chou KC (2008) Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. J Theor Biol 250:186–193

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17:729–738