Biochimie 95 (2013) 1741–1744

Contents lists available at SciVerse ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

Research paper

A protein structural class prediction method based on novel features

Lichao Zhang [a], Xiqiang Zhao [b,*], Liang Kong [c]

[a] College of Marine Life Science, Ocean University of China, Yushan Road, Qingdao 266003, PR China
[b] College of Mathematical Science, Ocean University of China, Songling Road, Qingdao 266100, PR China
[c] College of Mathematics and Information Technology, Hebei Normal University of Science and Technology, Qinhuangdao 066004, PR China

ARTICLE INFO

Article history:
Received 22 February 2013
Accepted 28 May 2013
Available online 14 June 2013

Keywords:
Sequence similarity
Protein structural class
Support vector machine
Secondary structure

ABSTRACT

In this study, a 12-dimensional feature vector is constructed to reflect the general contents and spatial arrangements of the secondary structural elements of a given protein sequence. Among the 12 features, 6 novel features are specially designed to improve the prediction accuracies for α/β and α + β classes based on the distributions of α-helices and β-strands and the characteristics of parallel β-sheets and anti-parallel β-sheets. To evaluate our method, the jackknife cross-validating test is employed on two widely-used datasets, 25PDB and 1189 datasets with sequence similarity lower than 40% and 25%, respectively. The performance of our method outperforms the recently reported methods in most cases, and the 6 newly-designed features have significant positive effect to the prediction accuracies, especially for α/β and α + β classes.

1. Introduction

The concept of protein structural class was introduced by Levitt and Chothia in 1976 [1]. In their study, 31 globular proteins were divided into four structural classes: all-α, all-β, α/β and α + β classes. It has been proved that information on the protein structural class plays important roles in many aspects of protein research. More specifically, a knowledge of structural classes has been applied to improve the accuracy of secondary structure prediction [2], to reduce the search space of possible conformations of the tertiary structure [3–5], and to implement a heuristic approach to determine tertiary structure [6]. The wide range of the applications of protein structural classes has been reviewed [7,8]. However, with the rapid development of the genomics and proteinics, the increasing gap between the output of sequencing and structural genomics creates difficulty in the advancement of research. Thus accurate and automated protein structural class prediction methods for newly-found proteins are urgently needed.

During the past three decades, a lot of computational methods have been developed for predicting the protein structural class. These methods mainly focused on sequence representation methods and classification algorithms. Several features based on the amino acid composition (AAC) [9–11] and pseudo amino acid composition (PseAAC) [12] have been applied to represent protein sequences. The prediction methods based on such features could achieve accuracies close to or more than 90% when tested on datasets of limited size or relatively high sequence similarity. However, by ignoring the predicted secondary structure information, these features performed poorly on datasets that were expanded or characterized by low-similarity, with accuracies between 50% and 70% [13]. Realizing the localization, many predicted secondary structure based features were introduced to improve the prediction accuracy [7,13–20], but the prediction accuracies for α/β and α + β classes are still unsatisfactory. This has become a deficiency in the current protein structural class prediction methods. After the features are extracted, various classification algorithms can be used to implement the protein structural class prediction, such as neural network, support vector machine (SVM), Bayesian classification, rough sets, fuzzy clustering, LogitBoost classifier and so on.

In this study, a new predicted secondary structure based method is proposed to predict protein structural class. First, a 12-dimensional feature vector is constructed. Among the 12 features, 6 novel features are specially designed to improve prediction accuracies, especially for α/β and α + β classes, and another 6 features have been used in the previous works. Then, the SVM classifier and jackknife test are adopted to predict and evaluate the model on two widely-used low-similarity benchmark datasets (25PDB and 1189 datasets). The results show that our method achieves satisfactory performance in comparison with other existing methods.

[*] Corresponding author. Tel.: +86 53266787282.
E-mail address: zhaodss@yahoo.com.cn (X. Zhao).

0300-9084/$ – see front matter © 2013 Elsevier Masson SAS. All rights reserved.
http://dx.doi.org/10.1016/j.biochi.2013.05.017

2. Materials and methods

2.1. Datasets

To evaluate the proposed method and facilitate its comparison with other existing methods, two widely-used benchmark datasets were adopted in our study. The 25PDB [21] and 1189 datasets [22] were selected from high-resolution protein structures, with low sequence similarity (no more than 25% and 40%, respectively). The 25PDB dataset contained 1673 protein domains, consisting of 443 all-α, 443 all-β, 346 α/β and 441 $\alpha + \beta$. The 1189 dataset included 1092 protein domains, of which 223 all-α, 294 all-β, 334 α/β and 241 $\alpha + \beta$.

2.2. Feature vector

Every amino acid residue in a protein sequence was predicted into one of the following three secondary structural elements: H (helix), E (strand) and C (coil). The predictions can be obtained from PSIPRED [23]. Since α-helices and β-strands were usually separated in α/β proteins, but were usually interspersed in $\alpha + \beta$ proteins, in order to reflect the distribution of α-helices and β-strands effectively, we constructed a simplified sequence from the primary predicted secondary structure sequence. For simplicity, the two sequences were abbreviated as SSS (secondary structure sequence) and SS (simplified sequence). First, every segment H, E and C in SSS was respectively replaced by the letters α, β and c. Then, all of the letters c were removed and SS was obtained. Here, the lengths of SSS and SS were denoted by N and N'. Based on SSS and SS, several features were rationally constructed to reflect the general contents and spatial arrangements of H, E and C. The details of these features were given as follows:

1. $p(\mathrm{H})$ and $p(\mathrm{E})$ expressed the fraction of the H and E in SSS. It had been proved that they were important to improve prediction accuracy [13].
2. The three features, $\mathrm{CMV_H}$, $\mathrm{CMV_E}$ and $\mathrm{CMV_C}$ [13] were proposed to reflect the spatial arrangement of H, E and C in SSS, respectively. They were formulated as:

$$\mathrm{CMV_H} = \frac{\sum_{j=1}^{N_\mathrm{H}} p_{\mathrm{H}_j}}{N(N-1)},\ \mathrm{CMV_E} = \frac{\sum_{j=1}^{N_\mathrm{E}} p_{\mathrm{E}_j}}{N(N-1)},\ \mathrm{CMV_C} = \frac{\sum_{j=1}^{N_\mathrm{C}} p_{\mathrm{C}_j}}{N(N-1)}$$

where N_H, N_E and N_C were the number of H, E and C residues in SSS, respectively; p_{H_j}, p_{E_j} and p_{C_j} were the jth position of H, E and C residues in SSS, respectively.

3. The 3-dimensional structure of protein was to some extent affected by the lengths of the structural elements such as α-helices and β-strands. Therefore, the normalized lengths of the longest α-helices and β-strands in SSS (denoted by $\mathrm{Maxseg_H/N}$ and $\mathrm{Maxseg_E/N}$) [17,18] were chosen in this study.

In order to better represent the level of separation and aggregation about α-helices and β-strands in SSS, the following novel features were specially designed for proteins from α/β and $\alpha + \beta$ classes.

4. The normalized maximum distances between the adjacent segment E and H ($\mathrm{Maxd_{EH}/N}$) as well as the adjacent segment H and E ($\mathrm{Maxd_{HE}/N}$) were proposed based on SSS.
5. The helix bundle probability ($P_{\alpha\alpha}$), sheet probability ($P_{\beta\beta}$) and crossing segments probability ($P_{\beta\alpha\beta}$) based on SS were introduced as follows:

$$P_{\alpha\alpha} = \frac{N_{\alpha\alpha}}{N'},\ P_{\beta\beta} = \frac{N_{\beta\beta}}{N'},\ P_{\beta\alpha\beta} = \frac{N_{\beta\alpha\beta}}{N'}$$

where $N_{\alpha\alpha}$, $N_{\beta\beta}$ and $N_{\beta\alpha\beta}$ were the number of the segments $\alpha\alpha$, $\beta\beta$ and $\beta\alpha\beta$, in SS.

6. Given that the β-strands were usually composed of parallel β-sheets in α/β protein and anti-parallel β-sheets in $\alpha + \beta$ protein, the distance probabilities ($P_{\mathrm{D_E}}$ and $P'_{\mathrm{D_E}}$) in SSS were defined as follows:

$$P_{D_\mathrm{E}} = \frac{N_1}{N_1 + N_2},\ P'_{D_\mathrm{E}} = \frac{N_2}{N_1 + N_2}$$

where D_E was the distance of the adjacent segment E, N_1 and N_2 were the number of the $D_\mathrm{E} \geq 5$ and $D_\mathrm{E} < 5$, respectively.

Among the above 14 features, 7 of them had been used in the previous works and the other features were newly designed to improve the prediction accuracies, particularly for the α/β and $\alpha + \beta$ classes. In order to improve the learning performance like prediction accuracy of the learning algorithm, the irrelevant and redundant features should be removed. Here, the wrapper model based feature selection algorithm was adopted to choose a subset of original features. It was performed on 25PDB dataset with SVM classifier as described in Section 2.3. As a result, a 12-dimensional structure feature vector (SFV) was selected and formally expressed as

$$\mathrm{SFV} = (p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, p_{12})^\mathrm{T}$$

where $p_1 = p(\mathrm{H})$, $p_2 = p(\mathrm{E})$, $p_3 = \mathrm{CMV_E}$, $p_4 = \mathrm{CMV_C}$, $p_5 = \mathrm{Maxseg_H/N}$, $p_6 = \mathrm{Maxseg_E/N}$, $p_7 = \mathrm{Maxd_{EH}/N}$, $p_8 = \mathrm{Maxd_{HE}/N}$, $p_9 = P_{\alpha\alpha}$, $p_{10} = P_{\beta\beta}$, $p_{11} = P_{\beta\alpha\beta}$ and $p_{12} = P_{D_\mathrm{E}}$. For example, given a secondary structure sequence SSS: EECEEECCEECCCCHHHHCCHHHCCCEEECCHHHCEE, its simplified sequence SS was $\beta\beta\beta\alpha\alpha\beta\alpha\beta$ with $N = 37$ and $N' = 8$. The corresponding 12-dimensional structure feature vector was

$$\mathrm{SFV} = \left(\frac{10}{37}, \frac{12}{37}, \frac{97}{666}, \frac{139}{666}, \frac{4}{37}, \frac{3}{37}, \frac{4}{37}, \frac{3}{37}, \frac{1}{8}, \frac{1}{4}, \frac{1}{8}, \frac{1}{2}\right)^\mathrm{T}.$$

2.3. Classification algorithm construction

The support vector machine (SVM) was an excellent machine learning algorithm which achieved superior classification performance compared with other algorithms [7]. There were generally four kernel functions to perform the prediction, which were linear function, polynomial function, sigmoid function and radial basis function (RBF). In this study, the RBF was adopted because of its superiority over other kernel functions [24,25]. The parameters C and γ were optimized based on 10-fold cross-validation on 25PDB dataset with a grid search strategy in the LIBSVM software [26,27], where $C \in [2^{-5}, 2^{15}]$ and $\gamma \in [2^{-15}, 2^{5}]$. Finally, the parameters $C = 8$ and $\gamma = 2$ were selected in our study.

2.4. Performance measures

In statistical prediction, the jackknife test was widely used to evaluate the performance of many predictors because of its rigour and objectivity [28,29]. Thus the jackknife test was employed in our study. For comprehensive evaluation, the individual sensitivity (or accuracy, denoted by Sens), the individual specificity (Spec) and Matthew's correlation coefficient (MCC) over each of the four structural classes, as well as the overall accuracy (OA) over the entire dataset were reported. These parameters were detailed as follows [30]: