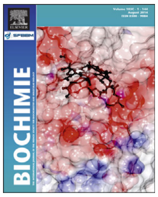
Biochimie 103 (2014) 131–136

Contents lists available at ScienceDirect

Biochimie

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



Research paper

A novel predictor for protein structural class based on integrated information of the secondary structure sequence

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

[a] Department of Biotechnology, College of Marine Life Science, Ocean University of China, Qingdao, PR China
[b] College of Mathematical Science, Ocean University of China, Qingdao, PR China
[c] College of Mathematics and Information Technology, Hebei Normal University of Science and Technology, Qinhuangdao, PR China

ARTICLE INFO

Article history:
Received 23 October 2013
Accepted 11 May 2014
Available online 22 May 2014

Keywords:
Sequence similarity
Protein structural class
Support vector machine
Position of α-helices and β-strands

ABSTRACT

The structural class has become one of the most important features for characterizing the overall folding type of a protein and played important roles in many aspects of protein research. At present, it is still a challenging problem to accurately predict protein structural class for low-similarity sequences. In this study, an 18-dimensional integrated feature vector is proposed by fusing the information about content and position of the predicted secondary structure elements. The consistently high accuracies of jackknife and 10-fold cross-validation tests on different low-similarity benchmark datasets show that the proposed method is reliable and stable. Comparison of our results with other methods demonstrates that our method is an effective computational tool for protein structural class prediction, especially for low-similarity sequences.

1. Introduction

The first definition of protein structural class was introduced by Levitt and Chothia in 1976 [1]. Based on their pioneering work, four structural classes of globular proteins are usually distinguished: (1) the all-α class, which includes proteins with only a small amount of strands, (2) the all-β class with proteins with only a small amount of helices, (3) the α/β class with proteins that include both helices and strands and where the strands are mostly parallel, and (4) the $\alpha + \beta$ class, which includes proteins with both helices and strands and where strands are mostly anti-parallel. The structural class has become one of the most important features for characterizing the overall folding type of a protein and plays important roles in many aspects of protein research. More specifically, a knowledge of structural class 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. To date, protein structural class prediction has become a quite meaningful topic in bioinformatics [6,7]. Traditional lab based methods assign the structural class to a protein by manual inspection such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, which is a time-consuming and expensive process. Thus, with the rapid development of the genomics and proteomics, it is crucially important to develop a fast and accurate computational method to determine structural class for the dramatically expanding newly-discovered proteins. One important aspect to predict structural class is to properly extract protein sequence information and then form a feature vector. In the earlier research, features were always extracted from amino acid (AA) sequence [4–13] such as the frequency of each AA in a given protein. Considering that these features ignored the sequential order information, some order features have been introduced, such as pseudo AA composition [14], collocation of AA, function domain composition [15], and position-specific scoring matrix (PSSM) computed by the position-specific iterative basic local alignment search tool (PSI-BLAST) [16]. However, these methods perform poorly with low-similarity sequences, with accuracies between 50% and 70% [17]. Recently, several new features based on predicted secondary structure sequence (SSS) have been proposed to improve the prediction accuracy with the low-similarity sequences [17–23] such as the length of the longest α-helices and β-strands. After the feature vector is extracted from the protein sequence, the feature vector is subsequently used as the input to different types of machine learning algorithms, including neural network [24], support vector machine (SVM) [25–29], fuzzy clustering [30], Bayesian classification [31], rough sets [32] and so on. A review by Chou

* Corresponding author. Department of Mathematics, College of Mathematical Science, Ocean University of China, Songling Road, Qingdao 266100, PR China. Tel.: +86 53266787282.
E-mail address: xqzhao62@163.com (X. Zhao).

http://dx.doi.org/10.1016/j.biochi.2014.05.008
0300-9084/© 2014 Elsevier Masson SAS. All rights reserved.

provided further details for the development of protein structural class prediction methods [6]. Although quite encouraging results have been achieved by many predicted secondary structure based methods, development of high quality prediction methods, especially for low-similarity sequences continues to be a challenging task.

In this study, an 18-dimensional integrated feature vector (IFV) is proposed by fusing the content and position information of the predicted secondary structure elements and then a multi-class support vector machine (SVM) is implemented to predict protein structural class on three different low-similarity benchmark datasets. In order to evaluate the proposed prediction method objectively, the jackknife cross-validation test and the 10-fold cross-validation test (10-CV) are implemented. Thanks to the comprehensive features which could represent enough protein sequence information to grasp the relationship between protein sequence and structural class, the experimental results demonstrate that our method is an effective computational tool for protein structural classes prediction. Moreover, the results suggest that further mining the integrated information about content and position based on the predicted secondary structure sequence is an effective way to improve prediction accuracy.

2. Materials and methods

2.1. Datasets

In order to give a comprehensive experimental comparison of different prediction algorithms, three widely-used benchmark datasets with low sequence identity were employed in our study. The ASTRAL dataset (including 7 classes) selected had sequence similarity lower than 20% and contains 6424 sequences [21]. Among the 7 classes, four major classes (all-α, all-β, α/β and $\alpha + \beta$) were selected in this study. The dataset with 5626 sequences was randomly divided into two equal subsets, one was used as the training set (ASTRAL$_{\text{training}}$) and the other was used as the test set (ASTRAL$_{\text{test}}$) [18]. The dataset 25PDB [10] that comprised 1673 proteins of about 25% sequence similarity. The final dataset, named 640 [18,22] comprised 640 proteins of about 25% sequence similarity. The details about the four datasets are shown in Table 1.

2.2. Feature vector

It was known that every residue in a protein sequence was predicted into one of three secondary structural elements H (helix), E (strand) and C (coil) using PSIPRED. These secondary structural elements defined the predicted secondary structure sequence (SSS) of a given protein. Based on the SSS, the following 27 features were given to identify protein structural class, including 11 reused features in the previous studies and 16 novel features. Below, we gave the concrete details and investigated how these features contributed to the prediction results.

1. $P(\mathrm{H})$ and $P(\mathrm{E})$ [17] based on the SSS which expressed the fraction of H and E can reflect the contents of H and E in the SSS. They were formulated as:

Table 1
The compositions of the datasets employed in our study.

Dataset	All-α	All-β	α/β	$\alpha + \beta$	Total
ASTRAL$_{\text{training}}$	640	662	748	763	2813
ASTRAL$_{\text{test}}$	640	662	747	764	2813
25PDB	443	443	346	441	1673
640	138	154	177	171	640

$$P(\mathrm{H}) = N_{\mathrm{H}}/N, \;\; P(\mathrm{E}) = N_{\mathrm{E}}/N$$

where N_{H} and N_{E} were the number of H, E in the SSS. The length of the SSS was denoted by N.

2. CMV$_{\mathrm{H}}$, CMV$_{\mathrm{E}}$ and CMV$_{\mathrm{C}}$ [17] based on the SSS have proved that they were useful for protein structural class prediction since they reflect the spatial arrangements of H, E and C in the SSS. They were formulated as:

$$\mathrm{CMV}_{\mathrm{H}} = \sum_{j=1}^{N_{\mathrm{H}}} P_{\mathrm{H}j} \Big/ (N(N-1))$$

$$\mathrm{CMV}_{\mathrm{E}} = \sum_{j=1}^{N_{\mathrm{E}}} P_{\mathrm{E}j} \Big/ (N(N-1))$$

$$\mathrm{CMV}_{\mathrm{C}} = \sum_{j=1}^{N_{\mathrm{C}}} P_{\mathrm{C}j} \Big/ (N(N-1))$$

where N_{C} was the number of C in the SSS, $P_{\mathrm{H}j}$, $P_{\mathrm{E}j}$ and $P_{\mathrm{C}j}$ were the jth position of H, E and C in the SSS.

3. As the concept of protein structural class was given according to globular protein, the lengths of the α-helices and β-strands can affect the spatial structure of protein. The normalized lengths of the longest α-helices and β-strands in the SSS [23] (denoted by MaxsegH/N and MaxsegE/N) were significant to improve the prediction accuracy.
4. If two segments of E are separated by segments of H, these two segments of E would tend to form parallel β-sheets. Otherwise, they would tend to form anti-parallel β-sheets. Take sequence EEEEECCHHHHHHCEEEECCCCHHHEEEECCCCEEEE as an example (Fig. 1), segment 1 and segment 2, as well as segment 2 and segment 3, are supposed to form parallel β-sheets, and segment 3 and segment 4 are supposed to form anti-parallel β-sheets. Consider that the β-strands in α/β proteins were usually composed of parallel β-sheets, while in $\alpha + \beta$ proteins the β-strands were usually composed of anti-parallel β-sheets, the number of β-strands (segments of E) that form parallel β-sheets and the number of β-strands that form anti-parallel β-sheets were important to identify α/β and $\alpha + \beta$ classes. Here the normalized parallel and anti-parallel β-sheets (Pn_{E}/N and APn_{E}/N) [20] were used in this study.
5. The normalized maximum distances between the adjacent segments E and H as well as the adjacent segments H and E (Maxd$_{\mathrm{EH}}$/N and Maxd$_{\mathrm{HE}}$/N) were used in this study.

The above 11 features were used due to their prior successful application in protein structural class prediction. Below, we give 16 novel features to improve the prediction accuracy, and hope that

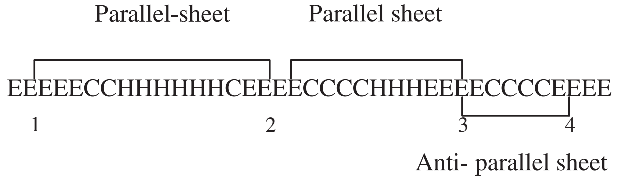


Fig. 1. The representation of E segments composing parallel β-sheets or anti-parallel β-sheets directly from the predicted secondary structural sequences.