Saudi Journal of Biological Sciences (2016) 23, 66–71

King Saud University

Saudi Journal of Biological Sciences

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

ORIGINAL ARTICLE

Prediction of complex super-secondary structure βαβ motifs based on combined features

Lixia Sun, Xiuzhen Hu [*], Shaobo Li, Zhuo Jiang, Kun Li

College of Sciences, Inner Mongolia University of Technology, Hohhot 010051, China

Received 11 September 2015; revised 8 October 2015; accepted 12 October 2015
Available online 12 November 2015

KEYWORDS

Structure prediction;
Complex super-secondary structure;
Combined features;
Support Vector Machine

Abstract Prediction of a complex super-secondary structure is a key step in the study of tertiary structures of proteins. The strand-loop-helix-loop-strand (βαβ) motif is an important complex super-secondary structure in proteins. Many functional sites and active sites often occur in polypeptides of βαβ motifs. Therefore, the accurate prediction of βαβ motifs is very important to recognizing protein tertiary structure and the study of protein function. In this study, the βαβ motif dataset was first constructed using the DSSP package. A statistical analysis was then performed on βαβ motifs and non-βαβ motifs. The target motif was selected, and the length of the loop-α-loop varies from 10 to 26 amino acids. The ideal fixed-length pattern comprised 32 amino acids. A Support Vector Machine algorithm was developed for predicting βαβ motifs by using the sequence information, the predicted structure and function information to express the sequence feature. The overall predictive accuracy of 5-fold cross-validation and independent test was 81.7% and 76.7%, respectively. The Matthew's correlation coefficient of the 5-fold cross-validation and independent test are 0.63 and 0.53, respectively. Results demonstrate that the proposed method is an effective approach for predicting βαβ motifs and can be used for structure and function studies of proteins.

1. Introduction

In proteins, if two secondary structure units are connected by a polypeptide (loop) with a specific arrangement of geometry, the resulting structure is referred to as a super-secondary structure or motif. Two or more super-secondary structures further fold into a complex super-secondary structure (Kuhn et al., 2004). The βαβ motif is a complex super-secondary structure in proteins (Yan and Sun, 1999), and it often appears in *Bacillus subtilis* proteases (Blundell et al., 1987). In strand-loop-helix-loop-strand structures, if there are one or more hydrogen bonds between two parallel β-strands, the structure is referred to as a βαβ motif.

The structure of a protein determines its function (Sun et al., 1997; Chou and Zhang, 1995; Chou, 1995). Thus, the prediction of protein structure is quite important in function research. At present, it is difficult to directly predict the tertiary structure from a protein sequence. Moreover, the super-secondary structure is a bridge between the secondary

[*] Corresponding author. Tel.: +86 471 6576281; fax: +86 471 6575863.
E-mail address: hxz@imut.edu.cn (X. Hu).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.sjbs.2015.10.005
1319-562X © 2015 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.

structure and tertiary structure, especially the complex super-secondary structure. Therefore, the prediction of complex super-secondary structure is a key step for the study of tertiary structure. With the increasing number of known protein structures and the well-developed feature selection algorithms such as mRMR (Peng et al., 2005), it is possible to develop theoretical methods to predict the complex super-secondary structure βαβ motifs in proteins.

The βαβ motif is an important complex super-secondary structure in proteins. In addition, many functional sites and active sites often occur in the polypeptides of βαβ motifs (Yan and Sun, 1999), including ADP-binding sites, FAD-binding sites, NAD-binding sites and other such functional sites (Wierenga et al., 1986). Therefore, the accurate prediction of βαβ motifs is very important to recognizing protein tertiary structure and the study of protein function.

Study of the βαβ motif began in 1983, where Taylor and Thornton correctly predicted the βαβ motifs with 70% accuracy in 16 α/β type proteins during predictions of a super-secondary structure (Taylor and Thornton, 1983). In 1984, they applied their method to identify 66 βαβ motifs in 18 proteins of α/β class with 75% accuracy (Taylor and Thornton, 1984). In 1986, Wierenga et al. predicted the occurrence of ADP-binding βαβ folds in the proteins from the PIR database, which contains 2676 proteins, by using amino acid sequence fingerprinting (Wierenga et al., 1986), but their dataset only involved the βαβ motif structure. Because the number of known protein structures was not sufficient at that time, they were limited to predicting the βαβ motif by using statistical methods. In more recent years, high-throughput technologies, information technology and computer technology have rapidly developed. The number of known protein structures has greatly increased, and it is feasible to predict the βαβ motif in a protein by using theoretical methods. In this study, we predicted the complex super-secondary structure βαβ motif, based on the principles and method which have been successfully used to predict β-hairpins of proteins in our previous work (Jia and Hu, 2011; Hu et al., 2010; Hu and Li, 2008).

The key step of complex super-secondary structure prediction is to construct a reasonable dataset and to select the best characteristic parameters and algorithm. In this paper, we constructed complex super-secondary structure βαβ motif datasets from protein structure data and used the sequence information, predicted structure information and function information to express the sequence features. To avoid an overfitting phenomenon when the higher dimension features are used in a Support Vector Machine algorithm, the dimensions of the amino acid component of position that is used to represent the sequence information were optimized by mRMR (Peng et al., 2005). Good predictive results were achieved according to 5-fold cross-validation and independent test.

2. Materials and methods

2.1. Materials

2.1.1. Dataset

In this paper, the βαβ motif datasets were constructed by using DSSP (Kabsch and Sander, 1983) in the following four steps:

(1) A dataset of 16,712 protein chains with $<95\%$ sequence identity was downloaded from ASTRAL 1.75 of SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/). We then deleted small proteins, and 14,977 protein chains were obtained. The dataset contained four classes of proteins: all α proteins, all β proteins, α/β proteins, and α + β proteins.

(2) BLAST software (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/2.2.6/) was used for obtaining proteins with $<25\%$ sequence identity. In total, 8704 protein chains were obtained. To accurately define the secondary structure in DSSP, 4442 protein chains were obtained whose length was more than 100 residues with a resolution <3.0 Å.

(3) Secondary structure was assigned to each amino acid of all the protein chains using DSSP. DSSP defines 8 states: H (α-helices), G (3_{10}-helices), I (π-helices), B (single β-bridge), E (β-ladder), T (hydrogen bonded turn), S (bend) and blank. However, prediction methods are normally assessed for only 3 states (H, C, and E), and therefore, the 8 states have to be reduced to 3 states. The secondary structures G, H and I were expressed by H. Both B and E were expressed by E, and the remaining states were expressed by C. Overall, 11,736 ECHCE patterns were obtained.

(4) In ECHCE patterns, if there were one or more hydrogen bonds between two parallel β-strands, then the structure was designated as a βαβ motif. Totally, there were 1635 protein chains which contained at least one βαβ motif. Overall, 4277 βαβ motifs and 3366 non-βαβ motifs were obtained.

2.1.2. The statistical analysis of the sequence segments

Loop-helix-loop (loop-α-loop) is a nucleation structure in the βαβ (β-loop-α-loop-β) motif. To choose the study objects, we performed a statistical analysis on the loop-α-loop structure. Results of this analysis are shown in Fig. 1.

Fig. 1 shows that the loop-α-loop lengths of the βαβ motifs and non-βαβ motifs are mainly concentrated between 10 and 26 amino acids which accounted for 85.6% of the total motifs. Therefore, we extracted the loop-α-loop that was 10–26 amino acids in length to study (Taylor and Thornton, 1984).

Furthermore, the statistical analysis of the segment length is performed and the results are shown in Table 1. The ideal fixed-length pattern was selected as 32 amino acids based on the average length of βαβ motifs and non-βαβ motifs, and Kumar's segment selection method in the prediction of β-hairpins (Kumar et al., 2005) (see the following section).

2.1.3. The selection of the fixed-length pattern

According to the statistical analysis of the loop-α-loop structure and the principles of Kuhn et al.(2004) and Hu and Li, 2008 used in β-hairpin prediction, two selection methods were generated. In the first method the beginning of the left loop was considered to be located at the fifth position. In the second method the end of the right loop was considered to be located at the twenty-eighth position in a fixed-length pattern. The two methods ensured that all loops can be included in the fixed-length pattern.

Based on the central alignment principles of Kumar's method in β-hairpin prediction (Kumar et al., 2005), the