Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Prediction of the parallel/antiparallel orientation of beta-strands using amino acid pairing preferences and support vector machines

Ning Zhang[a], Guangyou Duan[a], Shan Gao[a], Jishou Ruan[b], Tao Zhang[a,*]

[a] Key Laboratory of Bioactive Materials, Ministry of Education and the College of Life Sciences, Nankai University, PR China
[b] Chern Institute of Mathematics and College of Mathematical Sciences and LPKM, Nankai University, 300071, PR China

ARTICLE INFO

Article history:
Received 22 September 2009
Received in revised form
5 November 2009
Accepted 17 December 2009
Available online 24 December 2009

Keywords:
β-sheet
Protein structure
Encoding matrix
Strand-to-strand interactions
Machine learning

ABSTRACT

In principle, structural information of protein sequences with no detectable homology to a protein of known structure could be obtained by predicting the arrangement of their secondary structural elements. Although some *ab initio* methods for protein structure prediction have been reported, the long-range interactions required to accurately predict tertiary structures of β-sheet containing proteins are still difficult to simulate. To remedy this problem and facilitate *de novo* prediction of β-sheet containing protein structures, we developed a support vector machine (SVM) approach that classified parallel and antiparallel orientation of β-strands by using the information of interstrand amino acid pairing preferences. Based on a second-order statistics on the relative frequencies of each possible interstrand amino acid pair, we defined an average amino acid pairing encoding matrix (APEM) for encoding β-strands as input in the prediction model. As a result, a prediction accuracy of 86.89% and a Matthew's correlation coefficient value of 0.71 have been achieved through 7-fold cross-validation on a non-redundant protein dataset from PISCES. Although several issues still remain to be studied, the method presented here to some extent could indicate the important contribution of the amino acid pairs to the β-strand orientation, and provide a possible way to further be combined with other algorithms making a full 'identification' of β-strands.

1. Introduction

Current methods of protein tertiary structure prediction (e.g. Hidden-Markov models (HMM), sequence profile searches, and protein threading methods (Jones, 1999; Kelley et al., 2000; Schaffer et al., 1999; Shi et al., 2001)) typically use folds of known structures as templates. ROSETTA (Bonneau et al., 2001; Simons et al., 1997), one of the most successful approaches to tertiary structure prediction, generates a distribution of plausible local conformations for each segment of a chain by searching the PDB for fragments with similar local sequences. TASSER, developed by Zhang and Skolnick (2004), improved the prediction of protein tertiary structures that have lower sequence identity (especially on average 22%) to known structures. However, these methods are still less accurate for assigning a protein structure on the basis of the sequence alone when a protein has no close homologue in the protein data bank (Dou et al., 2004).

In principle, structural information for protein sequences with no detectable homology to a protein of known structure could be obtained by predicting the arrangement of their secondary structural elements (Steward and Thornton, 2002). As we know, the two predominant protein secondary structures are α-helices and β-sheets. Comparatively little is known about β-sheet structure (Jager et al., 2007). Although some ab initio methods for protein structure prediction have been reported (Bonneau et al., 2001; Bystroff et al., 2000; Ortiz et al., 1999; Osguthorpe, 2000; Samudrala et al., 1999; Simons et al., 1997), the long-range interactions required to accurately predict tertiary structures of β-sheet containing proteins are still difficult to simulate (Steward and Thornton, 2002). Therefore, β-sheet containing proteins have been particularly challenging for de novo structure prediction methods (Kuhn et al, 2004).

In a β-sheet, the individual extended polypeptide segments, called β-strands, can be arranged side by side to form a structure resembling a series of pleats. The β-strands can be parallel (N-termini of both strands at the same end) or antiparallel (otherwise) interacting with each other by hydrogen bonds. Earlier studies (Chou et al., 1983a, b, 1985, 1986, 1990; Chou and Carlacci, 1991; Chou and Scheraga, 1982) provide several classical computational works on parallel and antiparallel β-sheets. However, more and more new methods should be developed on newly constructed datasets to more deeply investigate β-sheet structures. In a β-sheet, adjacent β-strands bring distant residues into close contact with one another, and

* Corresponding author. Tel.: +86 022 23500237.
E-mail address: zhangtao@nankai.edu.cn (T. Zhang).

0022-5193/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2009.12.019

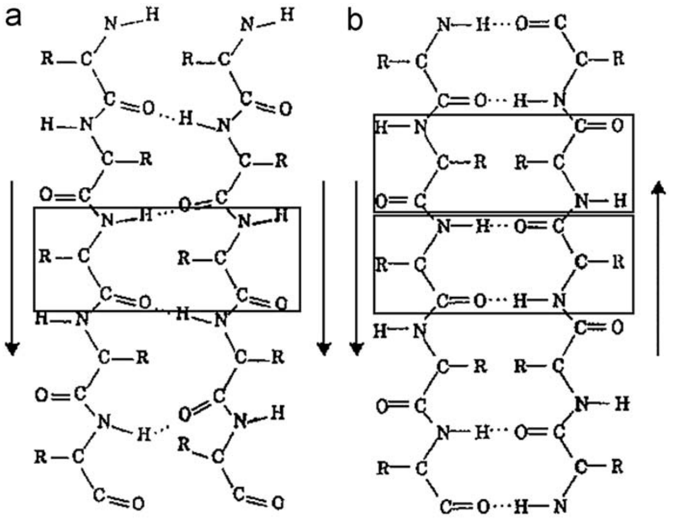

Fig. 1. A schematic diagram of (a) a parallel β-strand pair and (b) an antiparallel β-strand pair. The boxes around amino acids represent three interstrand amino acid pairs, respectively. Here, pairing amino acids are defined as those adjacent to each other in neighboring strands forming either 2 hydrogen bonds or 0 hydrogen bonds with each other (see Fooks et al., 2006). An amino acid pair in parallel β-strands has one HB (hydrogen-bonded) residue and one nHB (non-hydrogen-bonded) residue, while residues forming an interstrand pair within antiparallel strands are either both HB or both nHB.

constitute a specific mode of amino acid pairing (Fooks et al., 2006; Hutchinson et al., 1998; Lifson and Sander, 1980; Merkel et al., 1999; Wouters and Curmi, 1995) (like DNA base pairing) (Fig. 1). This mode of interaction among β-sheets occurs widely in protein tertiary structures, and plays an important role in protein folding pathways as well as in diseases ranging from AIDS and cancer to anthrax and Alzheimer's (Dou et al., 2004; Nowick, 2008).

Since assigning the β-strands topology of a β-sheet containing protein would reduce the three-dimensional space to be searched by ab initio methods (Kolinski et al., 2001; Steward and Thornton, 2002), there is a growing recognition of the importance of the strand-to-strand interactions among β-sheets (Nowick, 2008). Several studies, including statistical studies examining the frequencies of nearest-neighbor amino acids in β-ladders, found a significantly different preference for certain interstrand amino acid pairs at non-hydrogen-bonded and hydrogen-bonded sites (Cochran et al., 2001; Fooks et al., 2006; Hutchinson et al., 1998; Jager et al., 2007; Lifson and Sander, 1980; Merkel et al., 1999; Russell and Cochran, 2001; Wouters and Curmi, 1995). Dou et al (2004) created a comprehensive database of interchain β-sheet (ICBS) interactions. In our previous study, a database named SheetsPair (Zhang et al., 2007a) was also developed to compile both the interchain and the intrachain amino acid pairs.

At the most straightforward level, full 'identification' of a β-strand pair could consist of (i) finding the interacting partner β-strand(s), (ii) predicting the relative orientation (i.e. parallel or antiparallel) and (iii) shifting the relative positions of the two interacting β-strands. However, a cooperative folding event is often difficult to characterize. Studies ought to be performed separately on these concrete aspects. Previous efforts on (i) mainly focused on the predictions of residue contact map (Such as Zhang et al., 2005a, b; Cheng and Baldi, 2007) or amino acid partners in β-sheet (such as Baldi et al., 2000). Steward and Thornton (2002) developed an information theory approach to predict the relative offset positions by shifting one β-strand up to 10 residues either side of that observed (i.e. the aim (iii)). They treated parallel and antiparallel β-strands separately and suggested significant differences in the distributions of interstrand amino acid pairs between the two types of β-strands. However, they did not present a method by which the orientation (i.e. parallel or antiparallel) could be determined, which is another important step in the β-strand pair full 'identification'. In the present study, we developed a support vector machine-based approach for predicting whether a pair of β-strands adopts parallel or antiparallel orientation, i.e., trying to achieve the aim (ii) only.

2. Methods

2.1. Dataset

All protein data used in this study were taken from a PISCES (Wang and Dunbrack, 2003) dataset generated on May 16, 2009. In this dataset, the percentage identity cutoff is 25%, the resolution cutoff is 2.0 Å, and the *R*-factor cutoff is 0.25. All data were further preprocessed according to the following criterions: (i) no β-sheet containing protein chains were removed; (ii) protein chains having non-standard three-letter residue names (such as DPN, EFC, ABA, C5C, PLP,) were removed, since these indicate that the protein chains have covalently bounded ligands or modified residues; (iii) protein chains with uncertain structures or incorrect data was removed. Finally, 2,315 protein chains were extracted. In this study, we treat a pair of interacting β-strands as one sample. There are 6,786 parallel β-strand pairs (positive samples) and 12,734 antiparallel β-strand pairs (negative samples). In total, the dataset has 94,599 interstrand amino acid pairs on all β-strand pairs. We use 1 to stand for parallel β-strand pairs, while −1 to antiparallel ones. The dataset is available online and can be downloaded for academic use (http://sky.nankai.edu.cn/script/sky/english/bioinfo/PApreDS.zip).

2.2. Amino acid pairing preferences

Several statistical studies have been performed on the interstrand amino acid pairing preferences (Fooks et al., 2006; Hutchinson et al., 1998; Wouters and Curmi, 1995). In this study, a second-order statistics is made (Dou et al., 2004), and two relative frequency matrices are obtained for parallel (Table 1) and antiparallel (Table 2) β-strands, respectively. An element in each of the matrices is defined as following:

$$m_{ij} = P(A_i : A_j)/(P(A_i)P(A_j)), 1 \leq i \leq 20, 1 \leq j \leq 20, i \leq j$$

where m_{ij} is one element in the matrix. A_i and A_j are the two amino acids forming an interstrand pair. $P(A_i : A_j)$ represents the observed frequency of the amino acid pair $A_i : A_j$ on parallel or antiparallel β-strands. $P(A_i)$, $P(A_j)$ are the background probability of the amino acid A_i, A_j, respectively, by counting single amino acid frequencies across all protein sequences (not only β-structures) in the dataset. Thus, m_{ij} denotes a relative frequency (RF) of one possible amino acid pair formed by A_i and A_j. Note that the two matrices are both upper triangular matrices since we only consider 210 possible amino acid pairs, regardless of the order of the two amino acids within one pair. And we do not differentiate HB (hydrogen-bonded)/nHB (non-hydrogen-bonded) pairs or HB (hydrogen-bonded)/nHB (non-hydrogen-bonded) residues either (Fig. 1) (Nowick, 2008; Searle and Ciani, 2004).

The two relative frequency matrices (Tables 1 and 2) represent the amino acid pairing preferences on parallel and antiparallel β-strands, respectively. A *RF* with value > 1.0 indicates a favored