Journal of Theoretical Biology ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Identify five kinds of simple super-secondary structures with quadratic discriminant algorithm based on the chemical shifts

Gaoshan Kou, Yonge Feng*

College of Science, Inner Mongolia Agriculture University, Hohhot 010018, PR China

HIGHLIGHTS

- Chemical shift is used as feature for predicting protein super secondary structure.
- The quadratic discriminant analysis has been generalized to five groups.
- Predictive accuracy of CSs is superior to that of other feature by the same method.
- The results show chemical shift is an effective parameter in structure prediction.

ARTICLE INFO

Article history:
Received 9 April 2015
Received in revised form
2 June 2015
Accepted 4 June 2015

Keywords:
Chemical shifts
Analysis of variance
Quadratic discriminant analysis
Protein super secondary structure

ABSTRACT

The biological function of protein is largely determined by its spatial structure. The research on the relationship between structure and function is the basis of protein structure prediction. However, the prediction of super secondary structure is an important step in the prediction of protein spatial structure. Many algorithms have been proposed for the prediction of protein super secondary structure. However, the parameters used by these methods were primarily based on amino acid sequences. In this paper, we proposed a novel model for predicting five kinds of protein super secondary structures based on the chemical shifts ($C\alpha, C\beta, H$). Firstly, we analyzed the statistical distribution of chemical shifts of six nuclei in five kinds of protein super secondary structures by using the analysis of variance (*ANOVA*). Secondly, we used chemical shifts of six nuclei as features, and combined with quadratic discriminant analysis (*QDA*) to predict five kinds of protein super secondary structures. Finally, we achieved the averaged sensitivity, specificity and the overall accuracy of 81.8%, 95.19%, 82.91%, respectively in seven-fold cross-validation. Moreover, we have performed the prediction by combining the five different chemical shifts as features, the maximum overall accuracy up to 89.87% by using the $C\alpha, C\beta, H$ of H_α, which are clearly superior to that of the quadratic discriminant analysis (*QDA*) algorithm by using 20 amino acid compositions (*AAC*) as feature in the seven-fold cross-validation. These results demonstrated that chemical shifts ($C\alpha, C\beta, H$) are indeed an outstanding parameter for the prediction of five kinds of super secondary structures. In addition, we compared the prediction of the quadratic discriminant analysis (*QDA*) with that of support vector machine (*SVM*) by using the same six $C\alpha, C\beta, H$ as features. The result suggested that the quadratic discriminant analysis method by using chemical shifts as features is a good predictor for protein super secondary structures.

1. Introduction

Protein function is inherently correlated with its structure. Therefore, the study of protein structure is a basic premise for the prediction of its function. At present, it is still difficult to predict the spatial structure directly from amino acid sequence. However, the prediction of super secondary structure can contribute to predict protein tertiary structure, and it is a critical intermediate step toward the protein tertiary structure prediction. Protein super-secondary-structure motifs are composed of a few regular secondary structural elements connected by loops. Generally speaking, the empirical prediction of protein super secondary structure essentially consists of two parts: one is the prediction of different structural types from amino acid sequences (Burke and Deane, 2001; Bystro et al., 2000; Sun et al., 1997); another is the prediction of structural motifs (Chou, 1997a, 1997b; Chou, 2000a; Chou and Blinn, 1997). In this paper, we concentrate on the former. At present, there are five kinds of simple super secondary structures in ArchDB40 (Fernandez-Fuentes et al., 2004), namely, α-loop-α (*HH*),

Abbreviations: CSs, chemical shifts; AAC, amino acid compositions; ANOVA, analysis of variance; QDA, quadratic discriminant analysis; SVM, support vector machine; PseAAC, pseudoamino acid composition

* Corresponding author.
E-mail address: fengyonge@163.com (Y. Feng).

http://dx.doi.org/10.1016/j.jtbi.2015.06.006
0022-5193/

α-loop-β (*HE*), β-loop-α (*EH*), β-loop-β-hairpin (*EE*) and β-loop-β-link (*EE*1). These structural motifs play an important role in protein folding and stability, because a large number of motifs exist in protein spatial structure. Therefore, many research works have focused on proposing predictors for protein super secondary structure prediction (Blundell et al., 1988; Case, 1998; Chou, 2000a; Chou and Blinn, 1997; Cruz et al., 2002; Hu and Li, 2008). However, the features of these studies are mainly based on the amino acid compositions or dipeptide compositions. It is worth mentioning that Zou et al. (2011) used the approach of the pseudoamino acid composition (PseAAC) in predicting four kinds of simple super secondary structures, and achieved a good accuracy. The pseudoamino acid composition (PseAAC) (Chou, 2001, 2005) originally introduced by Prof. K.C. Chou was to avoid completely losing the sequence-order effects of a protein when it is represented by the conventional amino acid compositions (*AAC*). For a brief description about the original PseAAC, see a Wikipedia article at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition. Besides the *AAC* components, PseAAC also contained the 'pseudocomponents', through which the sequence-order effects of a protein are approximately reflected (Chou, 2000b; Shen and Chou, 2006, 2008). Later on the concept of PseAAC was extended to cover all the feature vectors of proteins (Chou, 2009, 2011). Furthermore, the concept of PseAAC has been extended to deal with DNA/RNA sequences (Chen et al., 2014; Lin et al., 2014; Liu et al., 2015b, 2015c). Because it has been widely and increasingly used in many areas of computational biology, recently a web server called 'Pse-in One' was established to generate various modes of pseudocomponents (Liu et al., 2015e), which is the first web server ever that can generate nearly all the features of pseudocomponents of DNA, RNA, and protein sequences in one package.

Nuclear magnetic resonance spectroscopy is a widely used technique in biochemistry that provides detailed information on the structure of molecules (Blundell et al., 1988; Chou, 2000a, 1997a, 1997b; Chou and Blinn, 1997; Cruz et al., 2002). However, chemical shift (Suzuki et al., 2014) can describe the local chemical environment of nuclear spins in nuclear magnetic resonance. Therefore, some researchers have utilized chemical shift for the determination of biomolecular structures (Case, 1998; Wishart and Case, 2001). Moreover, some works have studied on protein structure prediction (Cavalli et al., 2007; Lin et al., 2012; Mao et al., 2013; Mechelke and Habeck, 2013; Mielke and Krishnan, 2003; Pastore and Saudek, 1990; Shen et al., 2008; Wang, 2004; Zhang et al., 2003) and protein backbone and side chain torsion angle prediction (Shen and Bax, 2013) by using chemical shifts, these results showed that chemical shift is a powerful parameter for the determination of protein structure information.

In this paper, we would like to utilize chemical shifts (*CSs*) of nuclei as the parameters and combine with the method of quadratic discriminant analysis (*QDA*) to predict the five kinds of simple super secondary structures. Using the benchmark dataset, we adopted seven-fold cross-validation and obtained the averaged sensitivity, specificity and overall prediction accuracy of 81.8%, 95.19% and 82.91%, respectively by using six *CSs* as features. Moreover, we implemented the prediction by removing any one of the six nuclei and found that the chemical shift of each nuclei plays a different role in the prediction of protein super secondary structure. At the same time, in order to compare with other parameter, we have performed the prediction by using 20 amino acid compositions (*AAC*) as inputs of the method of quadratic discriminant analysis (*QDA*). The results showed that the performance of *CSs* outperforms that of 20 *AAC* in the five kinds of super secondary structures. In addition, we have performed the prediction by using the same six *CSs* as features of the method of support vector machine (*SVM*) in seven-fold cross-validation. Compared results showed that *QDA* is slightly better than *SVM* in terms of accuracies. As demonstrated by a series of recent publications (Chen et al., 2013; Ding et al., 2014; Xu et al., 2014; Chou, 2011), to establish a really useful sequence-based statistical predictor for a biological system, we should follow the following procedures: (a) construct or select a valid benchmark dataset to train and test the predictor; (b) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce or develop a powerful algorithm (or engine) to operate the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor. Below, let us describe how to deal with these steps one-by-one.

2. Materials and methods

2.1. Database

Firstly, chemical shifts of six nuclei $(C, C\alpha, C\beta, H, H\alpha, N)$ in proteins were selected from re-referenced protein chemical shift database (namely, RefDB; Zhang et al., 2003). Secondly, only the proteins were selected with super secondary structures information in ArchDB40 (Fernandez-Fuentes et al., 2004). Finally, the PISCES program (Wang and Dunbrack, 2005) was utilized to removing similar sequences. According to the aforementioned steps, 123 proteins were collected, which have both six *CSs* and super secondary structure. Among 123 proteins, all proteins have less than 30% sequence identity. Appendix A lists 123 proteins used in this paper. Finally, we got 110 α-loop-α (*HH*), 93 α-loop-β (*HE*), 110 β-loop-α (*EH*), 75 β-loop-β-hairpin (*EE*) and 157 β-loop-β-link (*EE*1) in the 123 proteins.

2.2. Feature parameter

It is one of the most important factors for pattern recognition to extract a set of informative parameters. Here, we used chemical shifts as features. In the five data subsets $\{HH, HE, EH, EE, EE1\}$, for a random sequence of length l, we calculated the averaged *CSs* of six nuclei $(C, C\alpha, C\beta, H, H\alpha, N)$ in the sequence by using the following Eq. (1):

$$t^m = \frac{1}{l}\sum_{j=1}^{l} CS_j^m \tag{1}$$

where CS_j^m denotes the chemical shift value of m nuclei ($m = C, C\alpha, C\beta, H, H\alpha, N$) for the jth residue in the sequence. Obviously, a sequence can be easily converted into a six-dimensional vector, called R: $\{t^m\}$.

2.3. Statistical distribution

The analysis of variance (*ANOVA*) can be used for multi-group samples means analysis of completely randomized design and provides a statistical test of whether or not the means of multi-group are all equal (Lin et al., 2012; Sprinthall, 2003). The difference of multi-group means can be measured by *ANOVA*, which is defined by Eq. (2)

$$MS_T = MS_B + MS_W \tag{2}$$

where MS_T, MS_B and MS_W denote the square means of total, between groups and within a group, respectively. The statistical value, called F-value, is the ratio of MS_B and MS_W, which can be calculated by Eq. (3)

$$F - \text{value} = MS_B / MS_W \tag{3}$$

From Eq. (3), we can see when the MS_B becomes increasingly larger than MS_W, F-value will become larger. That is to say, there are significant differences between groups, otherwise, the lack of differences.

Please cite this article as: Kou, G., Feng, Y., Identify five kinds of simple super-secondary structures with quadratic discriminant algorithm based on the chemical shifts. J. Theor. Biol. (2015), http://dx.doi.org/10.1016/j.jtbi.2015.06.006