Gene 642 (2018) 74–83

Contents lists available at ScienceDirect

Gene

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

Research paper

Protein secondary structure prediction based on the fuzzy support vector machine with the hyperplane optimization

Shangxin Xie[a], Zhong Li[a,*], Hailong Hu[a,b]

[a] School of Science, Zhejiang Sci-Tech University, Hangzhou, Zhejiang, 310018, China
[b] School of Science, Zhejiang A&F University, Lin'an, Zhejiang 311300, China

ABSTRACT

The prediction of the protein secondary structure is a crucial point in bioinformatics and related fields. In the last years, machine learning methods have become a valuable tool, achieving satisfactory results. However, the prediction accuracy needs to be further ameliorated. This paper proposes a new method based on an improved fuzzy support vector machine (FSVM) for the prediction of the secondary structure of proteins. Unlike traditional methods to set the membership function, it firstly constructs an approximate optimal separating hyperplane by iterating the class centers in the feature space. Then sample points close to this hyperplane are assigned with large membership values, while outliers with small membership values according to the K-nearest neighbor. And some sample points with low membership values are removed, reducing the training time and improving the prediction accuracy. To optimize the prediction results, our method also exploits information on sequence-based structural similarity. We used three databases (e.g. RS126, CB513 and data1199) to test this method, showing the achievement of 94.2%, 93.1%, 96.7% Q_3 accuracy and 91.7%, 89.7%, 94.1% SOV values for the three datasets, respectively. Overall, our method results are comparable to or often better than commonly used methods (Magnan & Baldi, 2014; Sheng et al., 2016) for secondary structure prediction.

1. Introduction

The prediction of protein secondary structure represents a crucial step for predicting the 3D structure of a protein and can also provide insight into protein function (Li et al., 2017). Currently, the continuous development of high-throughput instrumentations allowed gathering a large amount of protein sequence data. However, the achievement of protein structures starting from these data is still challenging, and there is an urgent requirement for using amino acid sequences to predict the protein structure and functions. The protein secondary structure prediction has thus become an important topic in bioinformatics and other related fields.

In this work, we present a new method based on an improved fuzzy support vector machine (FSVM) for the prediction of the protein secondary structure. To predict a secondary structure, classification rules referring to the description of each amino acid residue by using a series of discrete states are usually applied. There are different kinds of classification rules, and each one may have a different effect on the accuracy of the prediction (Kabsch & Sander, 1983; Richards & Kundrot, 1988). Here, we used the DSSP method (Kabsch & Sander, 1983) that assigns each amino acid to one of the three states (helix, sheet and coil). The assignment of several features to each amino acid, such as physical and chemical properties (Garnier et al., 1978), is the first step in the prediction of the secondary structure. We used the multiple sequence alignment tool Position Specific Iterative (PSI-BLAST) (Altschul et al., 1997) to generate features for each amino acid, using a 20-dimensional vector, and setting the size of the sliding window to 13 (Zheng et al., 2010).

Besides the extraction of amino acid features, choosing a proper prediction algorithm is another important step in the secondary structure prediction. Several methods have been proposed, such as those based on statistical approaches (Shen et al., 2005), but their performances were unsuccessful for large and complex biological datasets. Current methods are divided into two main categories, the template-based methods and the machine learning methods. Although the correlation between the sequence and structure similarity is not completely confirmed (Kaczanowski & Zielenkiewicz, 2010), it is well known that two domains with a similar sequence (segment) in general will show a similar structure (Lin et al., 2010). As a consequence, a high number of template-based prediction methods have been proposed

Abbreviations: SVM, Support vector machine; FSVM, Fuzzy support vector machine; Q_3, Overall accuracy; SOV, Segment overlap measure; MCC, Matthews correlation coefficient; PSI-BLAST, Position-Specific Iterative BLAST

[*] Corresponding author.
E-mail address: lizhong@zstu.edu.cn (Z. Li).

https://doi.org/10.1016/j.gene.2017.11.005
Received 8 September 2017; Received in revised form 29 October 2017; Accepted 2 November 2017

(Yaseen & Li, 2014), whose prediction accuracy for homologous proteins can reach about 80%. However, these algorithms are time consuming and strongly depend on the template database that researchers periodically have to check and update. For proteins whose templates are not included in the database, the correct prediction of the secondary structure is highly hampered. Therefore, template-based methods are not appropriate for non-homologous proteins (Bondugula & Xu, 2007). In the post genomic era and with the rapid increase of protein sequencing technologies, machine learning methods are undoubtedly valid to predict the protein secondary structure. Classical machine learning methods can achieve the accuracy of 80–83%, for example, by supporting vector machine (Suresh & Parthasarathy, 2014) and neural network (Qi et al., 2012). Among these machine learning methods, the support vector machine (SVM), firstly proposed by Cortes and Vapnik in 1995 (Cortes & Vapnik, 1995), is one of the most effective algorithms and has been widely used in the prediction of protein structure and function (Zhang et al., 2012). SVM maps the raw input data into a high-dimensional feature space, and then realizes the linear classification by constructing an optimal separating hyperplane that maximizes the margin between classes. The main advantage of SVM consists in the use of the Kernel function to solve the high-dimensional feature computation, ensuring the high generalization ability of the learning machine. Compared to the neural network, SVM can avoid the network structure selection and the local minimum problem. Recently, the application of membership functions for sample data has been introduced to form a fuzzy SVM (FSVM), which can improve the classification accuracy (Yang et al., 2011). Another trend in the prediction of protein secondary structure is to combine the machine learning methods with the comparison of sequence-based structural similarity according to the reference database (Magnan & Baldi, 2014), thus obtaining a prediction accuracy of 91–93%.

In this paper, we propose a new algorithm based on the enhanced fuzzy support vector machine. The improvement we achieved for the prediction accuracy of protein secondary structures is mainly attributed to: (a) a new membership function based on the approximate optimal separating hyperplane in the feature space is proposed. It guarantees that sample points near to the approximate optimal separating hyperplane, likely the support vectors, will be assigned with large membership values; otherwise, outliers and points apart from the approximate optimal separating hyperplane will be assigned with small membership values; (b) data in the training set are preprocessed before inputting to the FSVM, namely, samples with small membership values in the feature space will be removed to reduce the training time and improve the accuracy of the classification; (c) on the basis of Magnan and Baldi's method (Magnan & Baldi, 2014), proteins in the test set were subjected first to the comparison of sequence-based structural similarity with a reference protein database, and then to our FSVM method to predict the secondary structure.

2. The algorithm for the prediction of the protein secondary structure based on FSVM

2.1. A new fuzzy support vector machine

Recently, SVM has been widely used in bioinformatics and other related fields. However, SVM equally deals with the training samples when applied to predict the secondary structure, and sometimes can also cause the over-fitting problem (Cawley & Talbot, 2010). Unlike the traditional SVM, FSVM can emphasize the influence of the so-called support vectors and reduce the influence of the redundant training samples and outliers by setting the fuzzy membership value for each sample (Batuwita & Palade, 2010).

When using FSVM for the protein secondary structure prediction, the construction of an appropriate membership function is a key point. Classically, the fuzzy membership function is set according to the distance between the sample point and its class center (Zhang, 1999) in the input space. Before the classification of data, SVM needs to map the raw input data into a high-dimensional feature space. Therefore, the performance of FSVM based on these kinds of fuzzy memberships will be unsatisfactory. To address this problem, a new membership function calculated in the feature space has been proposed (Jiang et al., 2006) and defined as

$$\mathrm{s_i} = 1 - \sqrt{\mathrm{d_i^2}/(\mathrm{r_p^2} + \delta)} \tag{1}$$

where $\mathrm{r_p}$ is the radius of the class $\mathrm{C_p}$ defined as $\max\limits_{\mathrm{X_I \in C_P}} \|\Phi_\mathrm{p} - \Phi(\mathrm{X_i})\|$, $\mathrm{d_i}^2$ is the square of the distance between the training sample $\mathrm{X_i \in C_p}$ and its class center, which is defined as $\mathrm{d_i^2} = \mathrm{K(X_i, X_i)} - \frac{2}{\mathrm{n_p}} \sum_{\mathrm{X_{j \in C_p}}} \mathrm{K(X_i, X_j)} + \sum_{\mathrm{X_{j \in C_p}}} \sum_{\mathrm{X_{k \in C_p}}} \mathrm{K(X_j, X_k)}$, the corresponding class center of class $\mathrm{C_p}$ is defined as $\Phi_\mathrm{p} = \frac{1}{\mathrm{n_p}} \sum_{\mathrm{X_i \in C_P}} \Phi(\mathrm{X_i})$, where n_p is the number of samples in class $\mathrm{C_p}$, $\mathrm{K(X_i, X_j)} = \Phi(\mathrm{X_i})^\mathrm{T} \Phi(\mathrm{X_j})$ is a kernel function, and δ is a small number to avoid the case when $\mathrm{s_i}$ is equal to 0. All these expressions are calculated in the feature space.

Although this membership function is defined in the feature space, the generalization ability of the corresponding FSVM is actually not improved. According to this membership function, sample points away from the corresponding class center are assigned with relative small membership values (Fig. 1, blue points), while points near to their class center are assigned with large membership values (Fig. 1, red points). This assignment is not appropriate for the classification principle of SVM, because samples close to the corresponding class boundary compared with other samples are likely to be support vectors. Considering this problem and also that outliers, or noisy points (Fig. 1, green points) will easily lead to an over-fitting classification, we decided to introduce a new fuzzy membership function setting method based on an approximate optimal separating hyperplane, thus applying it to the FSVM for the prediction of the protein secondary structure.

The main idea of our method consists in a first setting of two initial hyperplanes which pass through the centers of two classes, followed by the construction of an approximate optimal hyperplane between the two initial hyperplanes, which can roughly separate sample points in the training data. The so-called approximate optimal separating hyperplane is constructed on the basis of the principle of SVM (maximizing the margin between two classes), namely, we use two iterative steps to construct an approximate separating hyperplane. We first determine two class centers according to the weighted distance keeping that the sum of the distances from each sample in the same class to the class center is the minimum. After fixing the hyperplane direction which is parallel to the line composed of two class centers, we translate

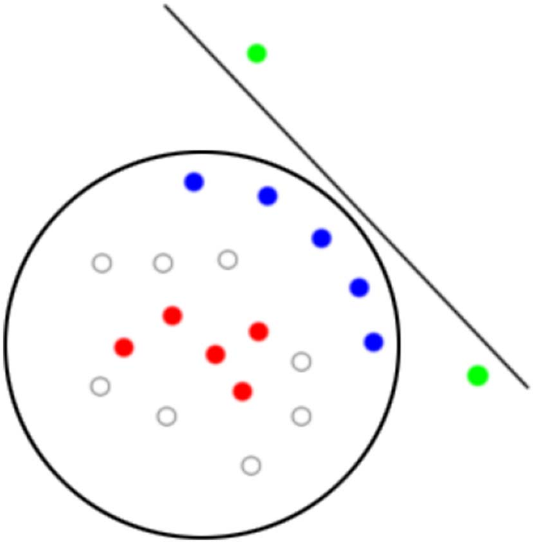

Fig. 1. Fuzzy membership based on the class center.