Journal of Theoretical Biology 400 (2016) 1–10

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Structural class prediction of protein using novel feature extraction method from chaos game representation of predicted secondary structure

Lichao Zhang [a,*], Liang Kong [b], Xiaodong Han [c], Jinfeng Lv [b]

[a] School of Mathematics and Statistics, Northeastern University at Qinhuangdao, Qinhuangdao 066004, PR China
[b] School of Mathematics and Information Science & Technology, Hebei Normal University of Science & Technology, Qinhuangdao 066004, PR China
[c] Zoucheng People's Hospital, Zoucheng 273500, PR China

HIGHLIGHTS

- Features are extracted from chaos game representation of secondary structure.
- Secondary structure distribution features can improve prediction significantly.
- Experimental results show that our feature extraction method is very promising.

ARTICLE INFO

Article history:
Received 11 January 2016
Received in revised form
18 March 2016
Accepted 8 April 2016
Available online 12 April 2016

Keywords:
Protein structural class
Secondary protein structure
Sequence similarity
Support vector machines

ABSTRACT

Protein structural class prediction plays an important role in protein structure and function analysis, drug design and many other biological applications. Extracting good representation from protein sequence is fundamental for this prediction task. In recent years, although several secondary structure based feature extraction strategies have been specially proposed for low-similarity protein sequences, the prediction accuracy still remains limited. To explore the potential of secondary structure information, this study proposed a novel feature extraction method from the chaos game representation of predicted secondary structure to mainly capture sequence order information and secondary structure segments distribution information in a given protein sequence. Several kinds of prediction accuracies obtained by the jackknife test are reported on three widely used low-similarity benchmark datasets (25PDB, 1189 and 640). Compared with the state-of-the-art prediction methods, the proposed method achieves the highest overall accuracies on all the three datasets. The experimental results confirm that the proposed feature extraction method is effective for accurate prediction of protein structural class. Moreover, it is anticipated that the proposed method could be extended to other graphical representations of protein sequence and be helpful in future research.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Protein structural class is an important property for characterizing the overall folding process of a protein and can provide important information about protein structure and function analysis, drug design and many other biomedical applications (Chou and Zhang, 1995; Zhou and Assa-Munt, 2001; Chou, 2004). According to the definition introduced by Levitt and Chothia (1976), a protein is generally categorized into one of the four structural classes, namely all-α, all-β, α/β and $\alpha+\beta$. The all-α and all-β classes represent structures that consist of mainly α-helices and β-strands, respectively. The α/β and $\alpha+\beta$ classes contain both α-helices and β-strands where the α/β class includes mainly parallel β-sheets and the $\alpha+\beta$ class includes anti-parallel β-sheets. These four structural classes also cover about 90% entries in the renowned manually annotated Structural Classification of Proteins (SCOP) database (Murzin et al., 1995) which has been regarded as the most accurate classification of protein structure. However, due to the rapid development of genomics and proteomics, the output of manually determined protein structural class has extremely lagged far behind the output of newly discovered protein sequences. It has become a huge barrier for further research.

With the development of bioinformatics, many computational methods have been proposed to automatically determine protein

* Corresponding author.
E-mail address: zhanglichaoouc@126.com (L. Zhang).

http://dx.doi.org/10.1016/j.jtbi.2016.04.011
0022-5193/© 2016 Elsevier Ltd. All rights reserved.

structural class (Zhang, 1994; Chou and Cai, 2004; Chou and Maggiora, 1998; Chou et al., 1998; Zhou, 1998; Chou, 2005; Xiao et al., 2008; Sahu and Panda, 2010; Kong et al., 2014; Zhang et al., 2014; Zhang, 2015; Li et al., 2014, 2015). These methods mainly differ in protein feature extraction and algorithm selection for classification, where more significant improvement for protein structural class prediction accuracy always comes from novel feature extraction methods (Dehzangi et al., 2014a). Numerous previous structural classes prediction methods extract features based on amino acid (AA) composition (Nakashima et al., 1986; Chou, 1995, 1999), which generally represents a protein as a 20-dimensional vector corresponding to the frequencies of 20 AAs in a given protein sequence. However, due to ignoring the important sequence order information, the corresponding prediction methods usually have mediocre performance. To overcome this limitation, some pseudo amino acid (PseAA) composition (Chou, 2001a; Xiao et al., 2006; Lin and Li, 2007) based features are proposed. Since these advanced features are sequence dependent, and since it has been conjectured that similar sequences share similar folding patterns, the corresponding prediction methods have been shown to be very successful in the prediction of protein structural classes for some high-similarity protein datasets (e.g., the ever commonly used 359 dataset, Chou and Maggiora, 1998, with sequence similarity over 95%). However, when the challenging low-similarity datasets (e.g., the 25PDB and 1189 datasets, Kurgan and Homaeian, 2006, with sequence similarities lower than 25% and 40%, respectively) are tested, these sequence-based features are not effective any more (Kurgan and Homaeian, 2006). Considering the fact that proteins with low sequence similarity but in the same structural class are likely to have high similarity in their corresponding secondary structure, several predicted secondary structure based feature extraction methods are proposed. Kurgan et al. (2008) compute the content of predicted secondary structural elements, count of segments, length of the longest segment, average length of the segment based on the predicted secondary structures in protein structural class prediction. In RKS-PPSC (Yang et al., 2010), 24 structural features are generated using recurrence quantification analysis, *k*-string based information entropy and segment-based analysis. Liu and Jia (2010) find that α-helices and β-strands alternate more frequently in α/β proteins than in $\alpha+\beta$ proteins, and count their alternating frequency as well as the content of parallel β-sheets and anti-parallel β-sheets. Zhang et al. (2011) compute the transition probability matrix of the reduced predicted secondary structural sequences and add it to protein structural class prediction. In PSCP-PSSE (Dai et al., 2013), some position-based statistical features of prediction secondary structural elements are proposed. Kong and Zhang (2014) extract comprehensive structure-driven features especially for segment distance-related features to predict protein structural class. With help of these structure-based features, the overall accuracy has been significantly enhanced on several low-similarity protein datasets. Despite the success of these predicted secondary structure based features, there is always space for improvement on the overall accuracy. Moreover, the predictions for the α/β and $\alpha+\beta$ classes are still of low quality when compared with the predictions for the all-α and all-β classes. These highlights the need for keeping on exploring novel feature extraction methods based on protein secondary structure.

The chaos game representation (CGR) is a graphical tool to provide intuitive picture for helping analyzing the complicated information hidden in a sequence (Niu et al., 2014). It was initially introduced by Jeffery to visualize DNA sequences (Jeffrey, 1990), and later applied to protein sequences as well (Basu et al., 1997; Yang et al., 2009, 2010; Niu et al., 2014; Fiser et al., 1994; Yu et al., 2004; Niu et al., 2012; Jingbo et al., 2011; Liu et al., 2011). Instead of performing feature extraction directly on the predicted secondary structure sequence, this study proposes a novel feature extraction method based on the CGR of predicted secondary structure to improve the prediction performance of protein structural class for low-similarity sequences. The general framework of the proposed method is shown in Fig. 1, which presents the pipeline that goes from the query sequence to the final output as well as intermediate steps. We first predict protein secondary structure sequence by PSIPRED program (Jones, 1999) and represent it as two time series by using CGR. Then, three groups of protein features are extracted to reflect content of secondary structure elements, sequence order information and secondary structure segments distribution information of a given protein sequence. The proposed feature vector is input to a multi-class nonlinear support vector machine (SVM) classifier to perform the prediction. Jackknife tests on three widely used low-similarity datasets (25PDB, 1189 and 640; Chen et al., 2008) and comparative analysis with state-of-the-art methods show the effectiveness of the proposed feature extraction method.

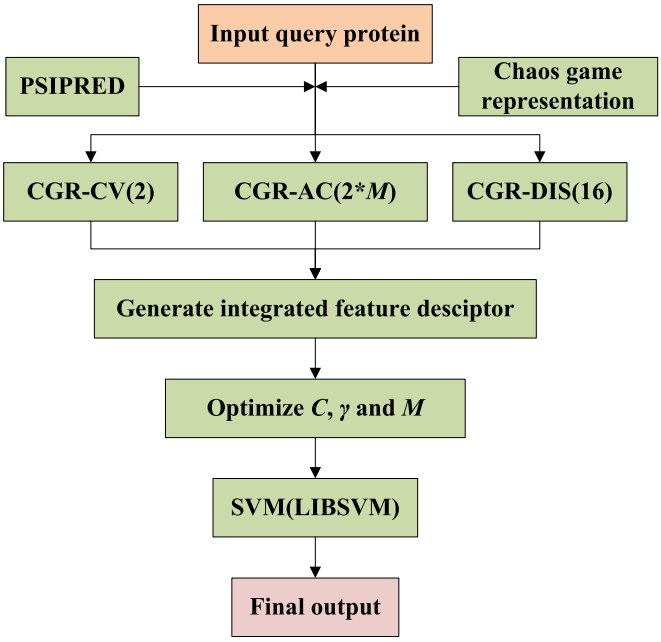

Fig. 1. The pipeline that goes from the query sequence to the final output as well as all intermediate steps.

2. Materials and methods

According to recent research (Kurgan and Homaeian, 2006), to establish a useful statistical predictor for a protein system, the following procedures should be considered: (1) selection of valid benchmark datasets to train and test the predictor, (2) representation of protein samples to reflect their intrinsic correlation with the target to be predicted, (3) selection of the classification algorithm to operate prediction, and (4) selection of the cross validation tests to objectively evaluate the anticipated accuracy of the predictor. Below, we will give concrete details about how to deal with these steps.

2.1. Datasets

Sequence similarity has a significant impact on prediction accuracy of protein structural class (Kurgan and Homaeian, 2006). In order to obtain more reliable prediction results and facilitate comparison with other existing methods, three widely used low-

Download English Version:

https://daneshyari.com/en/article/6369191

Download Persian Version:

https://daneshyari.com/article/6369191

Daneshyari.com