Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Structural bioinformatics

Improving Protein Fold Recognition by Extracting Fold-specific Features from Predicted Residue-residue Contacts

Jianwei Zhu [1,2], Haicang Zhang [1], Shuai Cheng Li [3], Chao Wang [1], Lupeng Kong [1, 2], Shiwei Sun [1], Wei-Mou Zheng [4,*] and Dongbo Bu [1,*]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
[4]Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Accurate recognition of protein fold types is a key step for template-based prediction of protein structures. The existing approaches to fold recognition mainly exploit the features derived from alignments of query protein against templates. These approaches have been shown to be successful for fold recognition at family level, but usually failed at superfamily/fold levels. To overcome this limitation, one of the key points is to explore more structurally-informative features of proteins. Although residue-residue contacts carry abundant structural information, how to thoroughly exploit these information for fold recognition still remains a challenge.

Results: In this study, we present an approach (called DeepFR) to improve fold recognition at superfamily/fold levels. The basic idea of our approach is to extract fold-specific features from predicted residue-residue contacts of proteins using deep convolutional neural network (DCNN) technique. Based on these fold-specific features, we calculated similarity between query protein and templates, and then assigned query protein with fold type of the most similar template. DCNN has showed excellent performance in image feature extraction and image recognition; the rational underlying the application of DCNN for fold recognition is that contact likelihood maps are essentially analogy to images, as they both display compositional hierarchy. Experimental results on the LINDAHL dataset suggest that even using the extracted fold-specific features alone, our approach achieved success rate comparable to the state-of-the-art approaches. When further combining these features with traditional alignment-related features, the success rate of our approach increased to 92.3%, 82.5%, and 78.8% at family, superfamily, and fold levels, respectively, which is about 18% higher than the state-of-the-art approach at fold level, 6% higher at superfamily level, and 1% higher at family level. An independent assessment on SCOP_TEST dataset showed consistent performance improvement, indicating robustness of our approach. Furthermore, bi-clustering results of the extracted features are compatible with fold hierarchy of proteins, implying that these features are fold-specific. Together, these results suggest that the features extracted from predicted contacts are orthogonal to alignment-related features, and the combination of them could greatly facilitate fold recognition at superfamily/fold levels and template-based prediction of protein structures.

Availability: Source code of DeepFR is freely available through https://github.com/zhujianwei31415/deepfr, and a web server is available through http://protein.ict.ac.cn/deepfr.

Contact: zheng@itp.ac.cn, dbu@ict.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Proteins are large molecules consisting of one or more chains of amino acid residues, and play crucial functions in a wide range of biological processes. The functional properties of proteins are largely determined by their three dimensional structures (called *tertiary structures*), making it vitally important to deduce or predict protein structures from amino acid sequences (Branden *et al.*, 1999). The widely-used experimental technologies to deduce protein structures, such as X-ray crystallography, NMR spectroscopy, and electron microscopy, have achieved great success; however, these technologies usually cost considerable time and thus cannot keep up with the fast process to acquire protein sequences (Berman *et al.*, 2000; Bairoch *et al.*, 2005). An alternative strategy is to predict protein structures from amino acid sequence, which can be divided into template-based (Roy *et al.*, 2010; Yang *et al.*, 2011; Ma *et al.*, 2012, 2014; Wang *et al.*, 2016) and *ab initio* prediction approaches (Simons *et al.*, 1997; Li *et al.*, 2008; Xu and Zhang, 2012).

During the whole procedure of protein structure prediction, recognizing the templates with similar structure to query protein, known as *fold recognition*, is an important and the first step (Jones *et al.*, 1992; Lindahl and Elofsson, 2000). The key point of fold recognition is to define and calculate the similarity between query protein and templates with known structures. The widely-used strategies for fold recognition include: (1) calculating alignment between query protein and templates based on sequence-sequence similarity (Altschul *et al.*, 1990; Pearson and Lipman, 1988; Eddy, 1998; Söding, 2005), or sequence-structure compatibility (also known as *threading*) (Shi *et al.*, 2001; Xu *et al.*, 2003; Peng and Xu, 2009); (2) designing a binary classifier to decide whether query protein is similar to templates based on specially-designed features (Cheng and Baldi, 2006; Dong *et al.*, 2009; Jo and Cheng, 2014; Jo *et al.*, 2015; Cheung *et al.*, 2016); and (3) adopting the consensus strategy to integrate multiple recognition approaches (Lundström *et al.*, 2001; Fischer, 2003; Xia *et al.*, 2016). These strategies mainly rely on sequence-related features and thus show excellent performance for fold recognition at family level. However, proteins in the same superfamily or fold usually show weak sequence similarities, making fold recognition at superfamily/fold levels a challenging task. This challenge emphasized the importance to exploit more structurally-informative features for fold recognition.

Residue-residue contacts present essential information of protein tertiary structures. Specifically, protein structures can be treated as result of the combination of local interactions and non-local interactions among residues, where local interactions lead to the formation of local structural motifs (e.g., secondary structural elements), and non-local interactions guide the arrangement of these motifs (Simons *et al.*, 1997; Li *et al.*, 2008; Zheng, 2014). According to the SCOP hierarchy, proteins in the same fold share similar arrangement or topological connections of secondary structural elements, and protein in the same superfamily share similar tertiary structures (Murzin *et al.*, 1995). Although the sequence similarity among these proteins might be relatively weak, the patterns of residue-residue contacts provide possibility to distinguish different superfamily or folds (Skwark *et al.*, 2014; Andreani and Söding, 2015).

This study presents such an effort to extract fold-specific features from predicted residue-residue contacts to improve fold recognition at superfamily/fold levels. Some ideas of this work come from two lines of research: predicting residue-residue contacts using evolutionary coupling analysis, and deep learning technique to extract fold-specific features and calculate similarity of contact likelihood matrices.

First, the multiple sequence alignment (MSA) of a query protein records its evolutionary history, where co-evolutionary events might occur among certain columns of MSA due mainly to contacts between corresponding residues. These co-evolutionary events, in turn, have been exploited to infer residue-residue contacts, generating contact likelihood maps as results (Göbel *et al.*, 1994; Olmea and Valencia, 1997). As transitive effects among interacting residues lead to considerable false-positive inference of contacts, direct evolutionary coupling (EC) analysis has been developed to reduce these transitive effects and improve contact prediction (Skolnick *et al.*, 1997; Morcos *et al.*, 2011; Jones *et al.*, 2012; Ekeberg *et al.*, 2013; Kamisetty *et al.*, 2013) and protein structure prediction (Kim *et al.*, 2014; Adhikari *et al.*, 2015; Wang *et al.*, 2017).

Second, deep convolutional neural network (DCNN) has proven to be successful for image retrieval (LeCun *et al.*, 1998; Krizhevsky *et al.*, 2012), and face recognition (Schroff *et al.*, 2015). One of the advantages of DCNN lies at its ability to automatically extract compositionally hierarchical features from images. For example, in images of certain objects, local features such as edges form motifs, motifs construct parts, and parts assemble into objects (LeCun *et al.*, 2015). Similar hierarchy also exists in protein structures and residue-residue contacts: under local interactions, residues form local structural motifs, and these local structural motifs are packed to form full structure under long-range interactions (Simons *et al.*, 1997; Haspel *et al.*, 2003; Taylor, 2016). This similarity implies the analog of contact maps of proteins to images of certain objects. Thus, the problem of inferring fold type for a protein from its contact map turns into the problem of image retrieval, making it reasonable to apply the DCNN technique for fold recognition. Recently, Wang *et al.*, 2017 successfully applied DCNN to remove noises from EC contact matrices, which also implies the possibility to apply DCNN for fold recognition.

It is worth noting that in previous studies, contact information has been explored for fold recognition in the form of contact order, contact number, and average contact probabilities of certain residue pairs in query protein (Cheng and Baldi, 2006). However, these approaches cannot thoroughly extract fold-specific information from contacts, which was confirmed by experiments to be shown in subsequent sections. Thus, a more effective way to explore fold-specific information from residue contacts was expected.

Combining these two lines of research, we proposed an approach called DeepFR that applies deep learning technique to extract fold-specific features from predicted contacts for fold recognition. The evaluation on two benchmark sets suggested that by using these extracted features, DeepFR outperformed the state-of-the-art approaches at superfamily/fold levels. Bi-clustering analysis validated that the features are fold-specific. The similarity of predicted contact maps between proteins are orthogonal to traditional features derived from alignments, and the combination of them could greatly facilitate accurate fold recognition as well as protein structure prediction.

2 Methods

The paradigm of our approach DeepFR is shown in Fig. 1, which consists of the following three steps: (1) *Contact prediction*: For a query protein, we first predicted its contact map through direct EC analysis of its MSA, yielding contact likelihood matrix as result; (2) *fold-specific features extraction*: As contact likelihood matrix usually contains a large amount of noises, we fed it into a pre-trained DCNN to extract fold-specific features, forming a EC-feature vector with fixed-size; (3) *Fold recognition*: Finally, the EC-feature vector of query protein was compared against those of templates with known fold type using a binary classifier, and the generated similarity scores were used to assign fold types for query protein. The details of these three steps are described as follows.