Computational Biology and Chemistry 48 (2014) 71–76

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Protein fold recognition based on functional domain composition☆

Qin Wang, Jinli Yan, Xiaoqin Li*

College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100124, People's Republic of China

ARTICLE INFO

Article history:
Received 16 September 2013
Accepted 9 December 2013

Keywords:
Functional domain composition
Fold recognition
Fold type
LIFCA database
Protein

ABSTRACT

Recognition of protein fold types is an important step in protein structure and function predictions and is also an important method in protein sequence-structure research. Protein fold type reflects the topological pattern of the structure's core. Now there are three methods of protein structure prediction, comparative modeling, fold recognition and de novo prediction. Since comparative modeling is limited by sequence similarity and there is too much workload in de novo prediction, fold recognition has the greatest potential. In order to improve recognition accuracy, a recognition method based on functional domain composition is proposed in this paper. This article focuses on the 124 fold types which have more than 2 samples in LIFCA database. We apply the functional domain composition to predict the fold types of a protein or a domain. In order to evaluate our method and its sensibility to the samples involving SCOP family divided, we tested our results from different aspects. The average sensitivity, specificity and Matthew's correlation coefficient (MCC) of the 124 fold types were found to be 94.58%, 99.96% and 0.91, respectively. Our results indicate that the functional domain composition method is a very promising method for protein fold recognition. And though based on simple classification rules, LIFCA database can grasp the functional features of different proteins, reflecting the corresponding relation between protein structure and function.

1. Introduction

Research of protein 3D structures plays a key role in molecular biology, cell biology, biomedicine, and drug design (Burley, 2000). With the technological improvements in protein crystal structure determination, especially in diversified structure determination methods, experimental determination could be carried out at a much faster speed. However, experimental determination still cannot keep pace with increasing protein sequences. Therefore, it is important to develop new methods to predict the 3D structure from amino acid sequences in the post-genome era. Currently, there are three methods of protein structure prediction: comparative modeling, fold recognition, and de novo prediction (Baker and Sali, 2001). Given that comparative modeling is limited by sequence similarity and there is too much workload in de novo prediction, fold recognition has the greatest potential in predicting protein structures.

Classification of protein fold type is a fundamental precondition in fold recognition. However, the prevailing classification database, such as SCOP (Murzin et al., 1995; Lo Conte et al., 2002) and CATH (Orengo et al., 1997; Pearl et al., 2005), have different classifications (Novotny et al., 2004; Matsuda et al., 2003) and are not constructed for fold recognition (Chen and Crippen, 2006). Thus, it is important to build a protein fold type database with a uniform principle for fold recognition research.

Achievements in fold recognition studies overseas have been reported. The fold recognition methods can be classified into the following three categories; based sequence (Dubchak et al., 1995, 1999; Ding and Dubchak, 2001; Shi et al., 2006; Jain et al., 2009), based structure (Marsolo, 2005; Marsolo and Parthasarathy, 2006), and fusion method of sequence and structure (Shi and Zhang, 2009; Shen and Chou, 2009; Gewehr et al., 2007; Ying et al., 2009). The methods are all based on the classification of SCOP, and involve less fold type. The 27 fold types of Ding are widely used. Support Vector Machines (SVM) was used to recognize the 27 fold types by Ding with a best success rate of 56.0% (Ding and Dubchak, 2001). Shi et al. used the SVM Fusion Network to predict fold types with an average accuracy of 61.04% (Shi et al., 2006), and they also used the image feature method, got an average accuracy of 71.95% (Shi and Zhang, 2009). Shen and Chou predicted the protein fold pattern with functional domain and sequential evolution information with a success rate of 70.5% for the 27 fold types (Shen and Chou, 2009). Liu et al. predicted the protein fold types by the general form of Chou's pseudo amino acid composition for the 27 fold types

☆ This work was supported by a grant from Natural Science Foundation of Beijing (4112010), Natural Science Foundation of China (31171267) and Beijing Municipal Commission (KM201110005027, KM201310005030).
* Corresponding author. Tel.: +86 010 67391610.
E-mail addresses: wangqin200819love@163.com, lxq0811@bjut.edu.cn (X. Li).

1476-9271/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2013.12.001

and obtained better identification results than most of the previous reported results (Liu et al., 2012). Ying et al. used a novel data integration approach to enhanced protein fold recognition for the 27 fold types and got a better result, the MKLdiv-dc method improved the fold discrimination accuracy to 75.19% (Ying et al., 2009). These researches play a key role in the method test, involving lesser samples in the database. In general, low accuracy was achieved for the fold recognition, based on small test sets.

Recently, some reports about large test sets for fold recognition achieved better results (Jain et al., 2009; Gewehr et al., 2007). Gewehr et al. used unique pattern-class mappings to make an automated prediction of SCOP classifications, where in the fold level, the average sensitivity was 93.36% and specificity was 98.13% (Gewehr et al., 2007). Jain et al. used supervised machine learning algorithms for protein structure classification based on SCOP classifications and found that the average sensitivities were 0.98, 0.75, 0.90, and 0.97 for the level of class, fold type, superfamily, and family, respectively (Jain et al., 2009).

Based on the purpose of build a protein fold type database with a uniform principle for fold recognition research, in earlier studies, a protein fold type database – LIFCA (low identical protein fold core structures and annotation) with a uniform principle according to the topological connection and spatial arrangement of secondary structure segments (α-Helix and β-Sheet) for protein folding type recognition was built (Luo and Li, 2000; Liu et al., 2008; Zhang et al., 2008), and good results have been achieved for the recognition of Globin-like fold (Ren et al., 2007) and another 36 large samples fold (Liu et al., 2009), here we defined the critical number is 4. The average sensitivity, specificity, and Matthew's correlation coefficient (MCC) of the 36 fold types were found to be 90.36%, 99.99% and 0.95, respectively. Results revealed that the HMM can be built for the less sampled fold type using the structure alignment tool together with manual inspection. The larger sample fold types which cannot be built with uniform HMM should be divided into subgroups so that the HMM can be built (Liu et al., 2009).

Recently, the functional domain composition method was widely used in bioinformatics, such as subcellular localization (Chou and Cai, 2004a), prediction of peptidase category (Xu et al., 2008), and protein structure class prediction (Chou and Cai, 2004b). Given that one protein or domain can contain one or more functional domains, and in general proteins which belong to the same fold type have similar functions, the functional domain composition method can also be used in fold recognition. In the present study, based on the purpose to test the new functional domain composition method in protein fold recognition, and also to test the classification of LIFCA which built in simple rules, we used this method in the chosen 124 fold types from LIFCA.

2. Materials and methods

2.1. Train set

With the principle of the topology invariance of protein structure' core, LIFCA database was built according to the topological connection and spatial arrangement of secondary structure segments. It contains 2406 proteins with less than 25% sequence identity, and there are 259 fold types. Compared with the SCOP database, LIFCA merged some SCOP families into one LIFCA fold, and also divided some SCOP families into different LIFCA folds.

A total of 124 fold types, which have more than 2 sequence samples were extracted from LIFCA. After removing the proteins without Pfam domain information, a training set with 2240 samples was obtained, covering 827 SCOP families. A total of 38 families were divided based on the LIFCA database.

2.2. Test set

To evaluate the current method and it is sensibility to the samples involving divided SCOP family, the current experimental results were tested from different aspects. A total of 9211 proteins with less than 95% sequence identity from the Astral 1.65 database were chosen as set A. The current method was also evaluated in a test set B, which excludes the duplicated proteins with the training set, and the proteins other than the 124 folds from the test set A, containing 2319 proteins, and 236 proteins involving divided SCOP families.

2.3. Method

The protein functional domain method was used as a predictor. First, the functional domains of the proteins in the training set were drawn by querying the Pfam database. As a result, the whole training set covers 1235 Pfam functional domains (Finn et al., 2006). Thus, each fold type and protein can be represented in the form of a 1235 dimension vector with each of the 1235 functional domains as the vector base. The feature vector F_i for a given fold type and the target vector P for a protein can be explicitly formulated as follows:

$$F_i = \begin{bmatrix} a_{i1} \\ a_{i2} \\ \cdots \\ a_{ij} \\ \cdots \\ a_{i1235} \end{bmatrix} \quad (i = 1\ldots124, j = 1\ldots1235); \quad P = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_j \\ \cdots \\ b_{1235} \end{bmatrix} \quad (j = 1\ldots1235);$$

where $a_{ij}, b_i = \begin{cases} 1, & \text{hit found} \\ 0, & \text{otherwise} \end{cases}$.

The similarity between P and F_i is described as follows:

$$\wedge(P, F_i) = \frac{P \cdot F_i}{||P|| \cdot ||F_i||} \quad (i = 1\ldots124)$$

where $P \cdot F_i$ is the dot product of P and F_i, $||P||$ and $||F_i||$ are their modules. Obviously, if P belongs to F_i, the similarity between them is highest. Accordingly, the predictor can be formulated as follows:

$$\wedge(P, F_k) = \text{Max}\{\wedge(P, F_1), \wedge(P, F_2), \ldots, \wedge(P, F_N)\}$$

If the similarity between P and F_k is the highest, P is predicted to belong to F_k.

2.4. Parameter estimation

The accuracy of the recognition results was estimated using Q, sensitivity, specificity, and MCC:

$$Q = \sum_{i=1}^{c} \frac{p_i}{N}$$

Sensitivity: $S_n = \frac{t_p}{t_p + f_n} \times 100\%$

Specificity: $S_p = \frac{t_n}{t_n + f_p} \times 100\%$

$$\text{MCC} = \frac{(t_p \times t_n) - (f_p \times f_n)}{\sqrt{(t_p + f_n) \times (t_n + f_p)} \times (t_p + f_p) \times (t_n + f_n)}$$

tp: true positive; tn: true negative; fp: false positive; fn: false negative.