Saudi Journal of Biological Sciences (2016) 23, 189–197

King Saud University

Saudi Journal of Biological Sciences

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

The recognition of multi-class protein folds by adding average chemical shifts of secondary structure elements

Zhenxing Feng [a], Xiuzhen Hu [a,*], Zhuo Jiang [a], Hangyu Song [a], Muhammad Aqeel Ashraf [b]

[a] Department of Sciences, Inner Mongolia University of Technology, Hohhot, China
[b] Water Research Unit, Faculty of Science and Natural Resources, University Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

Received 11 September 2015; revised 8 October 2015; accepted 12 October 2015
Available online 11 December 2015

KEYWORDS

Multi-class protein folds;
The increment of diversity;
Average chemical shifts;
Secondary structure elements;
Secondary structure motifs;
Random Forest algorithm

Abstract The recognition of protein folds is an important step in the prediction of protein structure and function. Recently, an increasing number of researchers have sought to improve the methods for protein fold recognition. Following the construction of a dataset consisting of 27 protein fold classes by Ding and Dubchak in 2001, prediction algorithms, parameters and the construction of new datasets have improved for the prediction of protein folds. In this study, we reorganized a dataset consisting of 76-fold classes constructed by Liu et al. and used the values of the increment of diversity, average chemical shifts of secondary structure elements and secondary structure motifs as feature parameters in the recognition of multi-class protein folds. With the combined feature vector as the input parameter for the Random Forests algorithm and ensemble classification strategy, we propose a novel method to identify the 76 protein fold classes. The overall accuracy of the test dataset using an independent test was 66.69%; when the training and test sets were combined, with 5-fold cross-validation, the overall accuracy was 73.43%. This method was further used to predict the test dataset and the corresponding structural classification of the first 27-protein fold class dataset, resulting in overall accuracies of 79.66% and 93.40%, respectively. Moreover, when the training set and test sets were combined, the accuracy using 5-fold cross-validation was 81.21%. Additionally, this approach resulted in improved prediction results using the 27-protein fold class dataset constructed by Ding and Dubchak.

* Corresponding author.
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

1. Introduction

The large numbers of protein sequences generated in the post-genomic era has challenged researchers to develop a high-throughput computational method to structurally

http://dx.doi.org/10.1016/j.sjbs.2015.10.008
1319-562X © 2015 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.

annotate these sequences. The protein fold reflects a key topological structure in proteins, as it contains three major aspects of protein structure: units of secondary structure, the relative arrangement of structures, and the overall relationship of protein peptide chains (Martin et al. 1998; Ming et al., 2015).

The proper spacial structure of a protein is highly correlated with its physiological functions. Abnormal protein folding may cause different diseases, for example, the neurodegenerative diseases such as Alzheimer's disease, spongiform encephalopathy, Parkinson's disease, mad cow disease etc. Thus, the correct identification of protein folds can be valuable for studies on pathogenic mechanisms and drug design (Thomas et al., 1995; Christopher and Michelle, 2004; Krishna and Grishin, 2005; Lindquist et al., 2001; Scheibel et al., 2004; Ma et al., 2002; Ma and Lindquist, 2002) and represents an important topic in bioinformatics.

In 2001, Ding and Dubchak (2001) constructed a dataset consisting of 27 protein fold classes using multiple feature parameters, including amino acid composition, predicted secondary structure, etc., and proposed support vector machines and neural network methods to predict the 27 protein fold classes, achieving an overall accuracy of 56.0%.

Subsequently, using the dataset constructed by Ding and Dubchak and identical feature parameters, several studies have suggested algorithmic improvements for protein fold identification. For example, Chinnasamy et al. (2005) introduced the phylogenetic tree and Bayes classifier for the identification of protein folds and achieved an overall accuracy of 58.2%. Nanni (2006) proposed a new ensemble of K-local hyperplanes based on random subspace and feature selection, achieving an overall accuracy of 61.1%. Guo and Gao (2008) presented a novel hierarchical ensemble classifier termed GAOEC (genetic-algorithm optimized ensemble classifier) and achieved an overall accuracy of 64.7%. Damoulas and Girolami (2008) proposed the kernel combination methodology for the prediction of protein folds and achieved an accuracy of 70%. Lin et al. (2013) exploited novel techniques to impressively increase the accuracy of protein fold classification.

Additional studies have suggested the selection of feature parameters to predict protein folds. For example, Shamim et al. (2007) used the structural properties of amino acid residues and amino acid residue pairs and achieved an overall accuracy of 65.2%. Dong et al. (2009) proposed a method termed ACCFold and achieved an overall accuracy of 70.1%. Nanni et al. (2010) proposed a method to extract features from the 3D structure and achieved significant improvement; however, this method does not solely rely on protein primary sequences to predict protein folds. Li et al. (2013) proposed a method termed PFP-RFSM and obtained improved results for protein fold identification.

Numerous studies have not only focused on the selection of feature parameters but also on the improvement of algorithms to identify protein folds. For example, Zhang et al. (2009) proposed an approach that utilizes the increment of diversity by selecting the pseudo amino acid composition, position weight matrix score, etc., and used these parameters to predict the 27 protein fold classes, with an overall accuracy of 61.1%. Shen and Chou (2006) applied the OET-KNN ensemble classifier to identify folds by introducing pseudo amino acids with sequential order information as a feature parameter and achieved an overall accuracy of 62.1%. Chen and Kurgan (2007) proposed the PFRES method using evolutionary information and predicted secondary structure, obtaining an accuracy of 68.4%. Ghanty and Pal (2009) proposed the fusion of heterogeneous classifiers approach, with features including the selected trio AACs and trio potential, and the overall recognition accuracy was 68.6%. Shen and Chou (2009) applied an identification method to protein folds using functional domain and sequential evolution information and achieved an overall accuracy of 70.5%. Yang and Kecman (2011) proposed a novel ensemble classifier termed MarFold, which combines three margin-based classifiers for protein fold recognition, and the overall prediction accuracy was 71.7%.

Additional studies have constructed and analyzed new 27-fold class datasets. For example, with a sequence identity less than 40%, Mohammad et al. (2007) constructed a dataset composed of 2554 proteins belonging to 27-fold classes, proposed structural properties of amino acid residues and amino acid residue pairs as parameters, and achieved an overall accuracy of 70.5% using 5-fold cross-validation. With sequence identity below 40%, Dong et al. (2009) constructed a 27-fold class dataset (containing 3202 sequences), proposed the ACC-Fold method, and obtained an overall accuracy of 87.6% using 5-fold cross-validation. Liu and Hu (2010) constructed a new 27-fold class dataset according to the construction of the Ding and Dubchak dataset (2001). This new dataset contains 1895 sequences with a sequence identity below 35%. Motif frequency, low-frequency power spectral density, amino acid composition, predicted secondary structure, and autocorrelation function values were combined as the set of feature parameters. Using the SVM algorithm and the ensemble classification strategy, the overall accuracy in the independent test was 66.67%. Moreover, studies on datasets consisting of 76, 86, and 199 fold classes have demonstrated improvements (Liu et al., 2012; Dong et al., 2009).

In this study, we reorganized the dataset constructed by Liu et al. (2012). According to the biological characteristics, values of the increment of diversity, motif frequency, predicted secondary structure motifs and the average chemical shift information of predicted secondary structure elements were extracted as feature parameters. Based on the ensemble classification strategy, these combined features were used as the input parameter for the Random Forests algorithm. An independent test and 5-fold cross-validation were used to predict the 76 protein fold classes, which resulted in good protein fold identification. The protein folds of the 27-fold class dataset and the corresponding structural classes were also identified, yielding improved results.

2. Materials and methods

2.1. Protein fold dataset

The 76-fold class dataset constructed by Liu et al. (2012) was reorganized; 8 and 5 protein sequences were added to the training and test set, respectively. Then the training set contains 1744 proteins for training, and the test set contains 1726 proteins for test. The sequence identity of the dataset was below 35%. The number of sequences of each type of protein fold was 10 or greater. The training and test set contained 1744 and 1727 protein chains, respectively. The distribution of the corresponding fold names and sequence numbers is shown in Table 1. The 76-fold class dataset is available at http://202.207.29.245:8080/Ha/HomePage/fzxHomePage.jsp.