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

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.