Available online at www.sciencedirect.com

SCIENCE DIRECT®

Biochemical and Biophysical Research Communications 334 (2005) 577–581

BBRC

www.elsevier.com/locate/ybbrc

Using supervised fuzzy clustering to predict protein structural classes

Hong-Bin Shen [a], Jie Yang [a], Xiao-Jun Liu [a], Kuo-Chen Chou [a,b,*]

[a] *Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030, China*
[b] *Gordon Life Science Institute, 13784 Torrey Del Mar, San Diego, CA 92130, USA*

Received 26 May 2005
Available online 1 July 2005

Abstract

Prediction of protein classification is both an important and a tempting topic in protein science. This is because of not only that the knowledge thus obtained can provide useful information about the overall structure of a query protein, but also that the practice itself can technically stimulate the development of novel predictors that may be straightforwardly applied to many other relevant areas. In this paper, a novel approach, the so-called "supervised fuzzy clustering approach" is introduced that is featured by utilizing the class label information during the training process. Based on such an approach, a set of "if-then" fuzzy rules for predicting the protein structural classes are extracted from a training dataset. It has been demonstrated through two different working datasets that the overall success prediction rates obtained by the supervised fuzzy clustering approach are all higher than those by the unsupervised fuzzy c-means introduced by the previous investigators [C.T. Zhang, K.C. Chou, G.M. Maggiora. Protein Eng. (1995) 8, 425–435]. It is anticipated that the current predictor may play an important complementary role to other existing predictors in this area to further strengthen the power in predicting the structural classes of proteins and their other characteristic attributes.

Keywords: Protein structural class; Fuzzy classifier; Supervised fuzzy classifier; Class label information; Jack-knife cross-validation test

The structural class is an important attribute used to characterize the overall folding type of a protein. Proteins often have quite similar or identical folding patterns even if they consist of very different sequences or bear various biological functions. In view of this, about 3 decades ago Levitt and Chothia tried to classify proteins into the following four structural classes: (1) all-α, (2) all-β, (3) α/β, and (4) $\alpha + \beta$. The all-α and all-β proteins are essentially formed by α-helices (Fig. 1A) and β-strands (Fig. 1B), respectively. The α/β class represents those proteins with both α-helices and β-strands that are largely interspersed in forming mainly parallel β-sheets (Fig. 1C), while the $\alpha + \beta$ class represents those also with both α-helices and β-strands but they are largely segregated in forming mainly antiparallel β-sheets (Fig. 1D). Prediction of protein structural class is an important topic in protein science (see, e.g., a review [1]). A series of previous studies have shown that some correlation between the protein structural class and amino acid composition does exist. Actually, many efforts were made to predict the structural classes of proteins based on their amino acid composition [1–11].

Zhang et al. [12] applied fuzzy clustering to predict protein structural classes. Results show that the fuzzy clustering approach yielded quite comparable results to other methods. Since the application of fuzzy clustering represented introducing a new branch of mathematics into molecular biology, it is worthwhile to conduct an in-depth investigation. In fact, the following two issues are worthy of our further consideration. (1) The fuzzy clustering algorithm in [12] was used to estimate the distribution of the training protein datasets, without utilizing the class label of each datum available for the identification in the training dataset. This is a kind of

---

[*] Corresponding author.
E-mail address: kchou@san.rr.com (K.-C. Chou).

0006-291X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2005.06.128

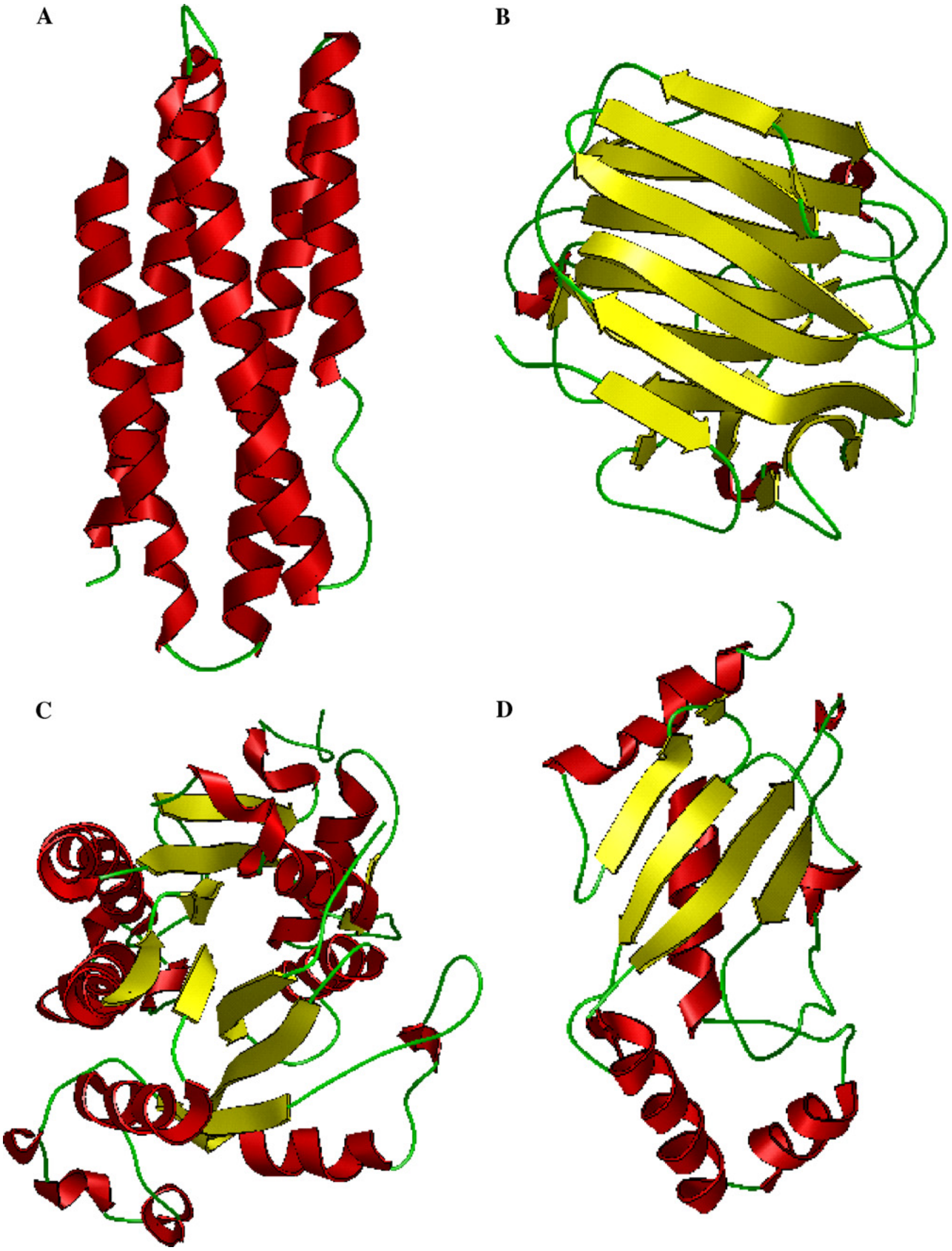

Fig. 1. Ribbon drawings to show the four structural classes of proteins: (A) all-α, (B) all-β, (C) α/β, and (D) α + β. Reproduced from [51] with permission.

"unsupervised" learning with fuzzy clustering that will loose the important label information. Actually, the class labels will provide useful guidance during the training process, such as in the applications of SVM, Neural Networks, and other similar learning machines. (2) As mentioned in [12], predicting the structural class of a given protein by the maximal membership function is relatively a simplistic approach; there are many other better options worthy of our consideration.

The present study was initiated in an attempt to introduce the supervised fuzzy clustering algorithm for predicting protein structural classes. The results thus obtained are better than those obtained by the unsupervised fuzzy clustering approach.

Methods

Fuzzy classifier. Typical fuzzy classifiers consist of interpretable if-then rules with fuzzy antecedents and class labels in the consequent part [13]. The antecedents (if-parts) of the rules partition the input space into a number of fuzzy regions by fuzzy sets, while the consequents (then-parts) describe the output of the classifier in these regions.