Predicting Protein Structural Class with AdaBoost Learner

Bing Niu[1,*], Yu-Dong Cai[1,2,4], Wen-Cong Lu[1], Guo-Zheng Li[3] and Kuo-Chen Chou[4]

[1]*Department of Chemistry, College of Sciences, Shanghai University, 99 Shang-Da Road, Shanghai 200436, China;* [2]*Department of Biomolecular Sciences, UMIST, Manchester M60 1QD, U.K;* [3]*School of Computer Science & Engineering, Shanghai University, Shanghai, 200072, China;* [4]*Gordon Life Science Institute, 13784 Torrey Del Mar, San Diego, CA 92130, USA*

Abstract: The structural class is an important feature in characterizing the overall topological folding type of a protein or the domains therein. Prediction of protein structural classification has attracted the attention and efforts from many investigators. In this paper a novel predictor, the AdaBoost Learner, was introduced to deal with this problem. The essence of the AdaBoost Learner is that a combination of many 'weak' learning algorithms, each performing just slightly better than a random guessing algorithm, will generate a 'strong' learning algorithm. Demonstration thru jackknife cross-validation on two working datasets constructed by previous investigators indicated that AdaBoost outperformed other predictors such as SVM (support vector machine), a powerful algorithm widely used in biological literatures. It has not escaped our notice that AdaBoost may hold a high potential for improving the quality in predicting the other protein features as well, such as subcellular location and receptor type, among many others. Or at the very least, it will play a complementary role to many of the existing algorithms in this regard.

Keywords: AdaBoost, component-coupled, neural network, SVM, domain structural classes, jackknife cross-validation.

1. INTRODUCTION

In spite of the fact that the details of the 3-dimensional structures of proteins and the domains therein are extremely complicated and irregular, their overall structural frames are surprisingly simple, regular [1,2], and truly elegant [3-7]. Many protein domains often have similar or identical folding patterns even if they are quite different according to their sequences [8-10]. Actually, such a phenomenon was observed by Levitt and Chothia about three decades ago on the basis of a quite limited number of structure-known proteins. They proposed to classify proteins into the following four structural classes: (1) all-α (Fig. **1a**) that is formed essentially by α-helices, (2) all-β (Fig. **1b**) essentially by β-strands, (3) α/β (Fig. **1c**) containing both α-helices and β-strands that are largely interspersed in forming mainly parallel β-sheets, and (4) α+β (Fig. **1d**) containing also both of the two secondary structure elements that, however, are largely segregated in forming mainly antiparallel β-sheets. The concept of structural class has ever since been widely used as an important attribute for characterizing the overall folding type of a protein or its domain (or domains).

Prediction of protein structural class is an important topic not only for the study of protein structure [11] but also for many other related areas. This is because it can not only provide useful information from the viewpoint of structure itself, but also greatly stimulate the characterization of many other features of proteins that may be closely correlated with

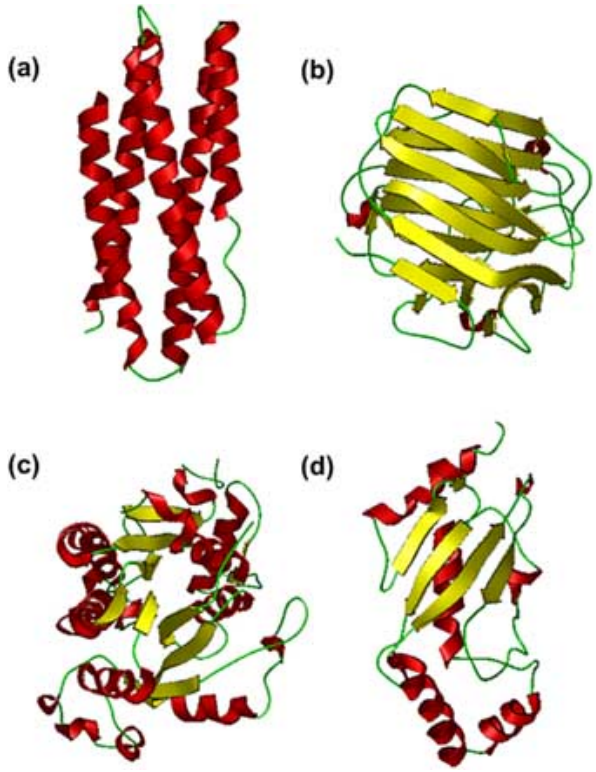

Figure 1. Ribbon drawings to show the four structural classes of proteins: (a) all-α, (b) all-β, (c) α/β and (d) α+β. Reproduced from [12] with permission.

*Address correspondence to this author at the Department of Chemistry, College of Sciences, Shanghai University, 99 Shang-Da Road, Shanghai 200436, China; E-mail: lifescience@san.rr.com

0929-8665/06 $50.00+.00

their biological functions [12]. Actually, many efforts have been made to predict the structural classes of proteins based on the knowledge of their sequences (see, e.g., [13-24]). Here we would like to introduce a complete different approach, the AdaBoost Learner, to deal with this problem. Because an individual domain is the most basic unit in structural classification [25-27], the present study will focus on protein domains.

2. METHODS

AdaBoost is a powerful method proposed originally for pattern recognition. The geometrical interpretation of AdaBoost is that it finds an accurate classification ruler, a strong learning algorithm, thru a combination with many other weak learning algorithms. The AdaBoost algorithm was proposed by Freund and Schapire [28]. They proved that an accurate 'strong' learning algorithm can be obtained by adding many 'weak' learning algorithms each of which performs just slightly better than a random guessing learning algorithm.

Suppose the input dataset is expressed by

$$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}, \quad y \quad \{-1, +1\}, \qquad (1)$$

where $\mathbf{x}_i$ represents the ith vector, and y_i the label of a target class. AdaBoost calls a given weak or base learning algorithm repeatedly in a series of time intervals $t = 1, 2, ..., T$. Each training sample is denoted with an equally initialized weight; i.e.,

$$\omega_{t,i} = \begin{cases} \dfrac{1}{2M}, & y_i = -1 \\ \\ \dfrac{1}{2L}, & y_i = 1 \end{cases} \qquad (2)$$

where L is the number of 'true' samples, and M is the number of 'false' samples. The training set was trained for T rounds. The target of training is to find an optimal classifier h_t and optimize it to be a strong classifier. It can be realized by increasing or decreasing the weights of classified examples after each training round to focus on the hard samples in the training set. The rule for updating the weight is given by

$$D_{t+1}(i) = \frac{D_t(i)\ e^{-\alpha_i y_i h_i(\mathbf{x}_i)}}{Z_t} = \frac{e^{-\sum_{j=1}^{t} \alpha_i y_i h_j(\mathbf{x}_i)}}{L \prod_{j=1} Z_j} = \frac{e^{-mrg(\mathbf{x}_i, y_i, f_i)}}{L \prod_{j=1} Z_j} \qquad (3)$$

where Z_t is a normalization factor, h_t is the base classifier, and α_i is a parameter that iteratively minimizes the potential of h_t, $mrg(\mathbf{x}_i, y_i, f_i)$ is functional margin of a point $(\mathbf{x}_i, y_i)$ with respect to the function.

$$Z_t = \sum_{i=1}^{L} D_t(i)\ \exp(-\alpha_i y_i h_t(\mathbf{x}_i)) \qquad (4)$$

where $D_t(i)$ is the weight of distribution on training example i at round [29]. Thus, the final classifier H can be obtained by combining many base classifiers thru the weighted majority vote; i.e.,

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_i h_t(x)\right) \qquad (5)$$

The skeleton of AdaBoost can be outlined as follows [29]: (1) establish $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), y_i \quad \{-1, +1\}$; (2) initialize and normalize the weight; (3) repeatedly execute the following steps from $t = 1, 2, ..., T$; (4) train the training set with the distribution D_t; (5) get the base classifier h_t which results in the least error; (6) update the weight focused on incorrect sample and set the new weight, eventually generating the final strong classifier H.

The AdaBoost algorithm was obtained from the software package Spider at http://www.kyb.tuebingen.mpg.de/bs/people/spider/. The AdaBoost method usually applies to two-class problems. For the current case of four-class problems, the "one-vs.-others" strategy [30] was adopted.

Choosing a proper weak classifier is important in performing AdaBoost algorithm. The following two criterions were followed to generate a good weak classifier: (1) the weak classifier should have the capacity of coping with reweighing of the data, and (2) the weak classifier should not result in over-fitting. Also, the KNN (K Nearest Neighbor) method was selected as the weak learning machine (weak classifier). The parameter K, i.e., the number of the nearest neighbors, was set at 1. After being trained, the thumb of rules output by the AdaBoost method was obtained, indicating that the trained model could be used for performing the prediction concerned.

3. RESULTS AND DISCUSSION

The datasets studied here were taken from Zhou [21]. The first dataset contains 277 protein domains, of which 70 are all-α domains, 61 all-β, 81 α/β, and 65 α+β. The second dataset contains 498 domains, of which 107 are all-α domains, 126 all-β, 136 α/β, and 129 α+β. The amino acid composition was used to represent the sample of a protein domain [14]. Therefore, each input of the AdaBoost actually corresponds to a vector in a 20-dimensional space [15,31].

To demonstrate the power of AdaBoost Learner, computations were performed by the jackknife cross-validation, which is deemed the most rigorous and objective test procedure in statistical prediction [17], and are widely used by more and more investigators in testing the power of various predictors [20,21,32-46].

The success rates thus obtained are given in Table **1**, where, for facilitating comparison, the corresponding rates

Table 1. Jackknife Cross-Validation Success Rates by Component-Coupled, Neural Network, SVM, and AdaBoost

Dataset	Algorithm	all-α	all-β	α/β	α+β	Overall
277 domains[a]	Component-coupled	84.3%	82.0%	81.5%	67.7%	79.1%
	Neural networks	68.6%	85.2%	86.4%	56.9%	74.7%
	SVMs	74.3%	82.0%	87.7%	72.3%	79.4%
	AdaBoost	87.1%	95.1%	98.7%	81.5%	90.9%
498 domains[b]	Component-coupled	93.5%	88.9%	90.4%	84.5%	89.2%
	Neural networks	86.0%	96.0%	88.2%	86.0%	89.2%
	SVMs	88.8%	95.2%	96.3%	91.5%	93.2%
	AdaBoost	96.2%	92.1%	98.5%	89.9%	94.2%

[a]Taken from Table I of Zhou [21].
[b]Taken from Table II of Zhou [21].

obtained by component-coupled algorithm [21,47], neural networks [48], and support vector machines (SVM) [49,50] are also listed. As we can see from the table, the current AdaBoost Learner is superior to all the other predictors in identifying the structural classification for the dataset of the 277 protein domains as well as the dataset of 498 domains.

4. CONCLUSION

The AdaBoost Learner is a very useful classifier. It has remarkably outperformed the powerful neural network and SVM classifiers in predicting the protein domain structural classes for the two datasets constructed by previous investigators. It is anticipated that the AdaBoost classifier can also be used to predict other protein attributes, such as subcellular localization [32,33,38,47,51-53], membrane types [40,41,54-57], enzyme family and subfamily classes [58-61], enzyme active sites [62,63], G-protein coupled receptor classification [64-66], and protein quaternary structure types [67], among many others.

REFERENCES

[1] Finkelstein, A. V. and Ptitsyn, O. B. (**1987**) *Prog. Biophys. Mol. Biol.*, *50*, 171-190.
[2] Chou, K. C. and Carlacci, L. (**1991**) *Proteins: Structure, Function, and Genetics*, *9*, 280-295.
[3] Chou, K. C. (**2004**) *Biochem. Biophys. Res. Commun.*, *316*, 636-642.
[4] Oxenoid, K. and Chou, J. J. (**2005**) *Proc. Natl. Acad. Sci. USA*, *102*, 10870-10875.
[5] Chou, K. C. (**2005**) *J. Proteome Res.*, *4*, 1681-1686.
[6] Chou, K. C. (**2004**) *Biochem. Biophys. Res. Commun.*, *319*, 433-438.
[7] Chou, K. C. (**1992**) *J. Mol. Biol.*, *223*, 509-517.
[8] Richardson, J. S. (**1977**) *Nature*, *268*, 495-500.
[9] Ptitsyn, O. B. and Finkelstein, A. V. (**1980**) *Quart. Rev. Biophys.*, *13*, 339-386.
[10] Chou, K. C. (**2004**) *Curr. Med. Chem.*, *11*, 2105-2134.
[11] Chou, K. C. (**2000**) *Curr. Protein and Peptide Science*, *1*, 171-208.
[12] Chou, K. C. (**2005**) *Curr. Protein and Peptide Science*, *6*, 423-436.
[13] Klein, P. and Delisi, C. (**1986**) *Biopolymers*, *25*, 1659-1672.
[14] Chou, J. J. and Zhang, C. T. (**1993**) *J. Theoretical Biology*, *161*, 251-262.
[15] Chou, K. C. and Zhang, C. T. (**1994**) *J. Biological Chem.*, *269*, 22014-22020.
[16] Mao, B., Chou, K. C. and Zhang, C. T. (**1994**) *Prot. Engineer.*, *7*, 319-330.
[17] Chou, K. C. and Zhang, C. T. (**1995**) *Crit. Rev. Biochem. Mol. Biol.*, *30*, 275-349.
[18] Bahar, I., Atilgan, A. R., Jernigan, R. L., and Erman, B. (**1997**) *Proteins: Structure, Function, and Genetics*, *29*, 172-185.
[19] Liu, W. and Chou, K. C. (**1998**) *J. Protein Chem.*, *17*, 209-217.
[20] Zhou, G. P. and Assa-Munt, N. (**2001**) *Proteins: Structure, Function, and Genetics*, *44*, 57-59.
[21] Zhou, G. P. (**1998**) *J. Protein Chem.*, *17*, 729-738.
[22] Cai, Y. D., Li, Y. X., and Chou, K. C. (**2000**) *BBA*, *1476*, 1-2.
[23] Cai, Y. D. and Zhou, G. P. (**2000**) *Biochimie*, *82*, 783-785.
[24] Chou, K. C. and Cai, Y. D. (**2004**) *Biochemical and Biophysical Research Communications* (Corrigendum: ibid., 2005, Vol. 329, 1362), 321, 1007-1009.
[25] Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (**1995**) *J. Molecular Biology*, *247*, 536-540.
[26] Chou, K. C., Liu, W., Maggiora, G. M., and Zhang, C. T. (**1998**) *PROTEINS: Structure, Function, and Genetics*, *31*, 97-103.
[27] Chou, K. C. and Maggiora, G. M. (**1998**) *Protein Engineering*, *11*, 523-538.
[28] Freund, Y. and Schapire, R. (**1997**) *J. Computer and System Sciences*, *55*, 119-139.
[29] Schapire, R. E. and Singer, Y. (**1999**) *Machine Learning*, *37*, 297-336.
[30] Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Ares, J. M., and Haussler, D. (**2000**) *Proc. Natl. Acad. Sci. USA*, *97*, 262-267.
[31] Chou, K. C. (**1995**) *Proteins: Structure, Function & Genetics*, 21, 319-344.
[32] Chou, K. C. (**2001**) *Proteins: Structure, Function, and Genetics* (Erratum: ibid., 2001, Vol.44, 60), 43, 246-255.
[33] Zhou, G. P. and Doctor, K. (**2003**) *Proteins: Structure, Function, and Genetics*, *50*, 44-48.
[34] Feng, Z. P. (**2001**) *Biopolymers*, *58*, 491-499.
[35] Luo, R. Y., Feng, Z. P., and Liu, J. K. (**2002**) *Eur. J. Biochem.*, *269*, 4219-4225.
[36] Chou, K. C. (**2000**) *Analytical BioChem.*, *286*, 1-16.
[37] Chou, K. C. and Cai, Y. D. (**2002**) *J. Biological Chem.*, *277*, 45765-45769.
[38] Xiao, X., Shao, S., Ding, Y., Huang, Z., Huang, Y., and Chou, K. C. (**2005**) *Amino Acids*, *28*, 57-61.
[39] Gao, Y., Shao, S. H., Xiao, X., Ding, Y. S., Huang, Y. S., Huang, Z. D., and Chou, K. C. (**2005**) *Amino Acids*, *28*, 373-376.
[40] Wang, M., Yang, J., Xu, Z. J., and Chou, K. C. (**2005**) *J. Theoretical Biology*, *232*, 7-15.
[41] Wang, M., Yang, J., Liu, G. P., Xu, Z. J., and Chou, K. C. (**2004**) *Protein Engineering, Design and Selection*, *17*, 509-516.

[42] Shen, H. P., Yang, J., Liu, X. J., and Chou, K. C. (**2005**) *Biochem. Biophys. Res. Commun.*, *334*, 577-581.
[43] Shen, H. and Chou, K. C. (**2005**) *Biochem. Biophys. Res. Commun.*, 288-292.
[44] Shen, H. B. and Chou, K. C. (**2005**) *Biochem. Biophys. Res. Comm.*, *337*, 752-756.
[45] Feng, K. Y., Cai, Y. D., and Chou, K. C. (**2005**) *Biochem. Biophys. Res. Commun.*, *334*, 213-217.
[46] Liu, H., Wang, M., and Chou, K. C. (**2005**) *Biochem Biophys Res Commun*, *336*, 737-739.
[47] Chou, K. C. and Elrod, D. W. (**1999**) *Prot. Engineering*, *12*, 107-118.
[48] Bishop, C. Neural Networks for Pattern Recognition; Oxford Press, **1995**.
[49] Cristianini, N. and Shawe-Taylor, J. *Support Vector Machines*; Cambridge University Press: Cambridge, **2000**.
[50] Vapnik, V. *Statistical Learning Theory*; Wiley-Interscience, New York, **1998**.
[51] Pan, Y. X., Zhang, Z. Z., Guo, Z. M., Feng, G. Y., Huang, Z. D., and He, L. (**2003**) *J. Protein Chem.*, *22*, 395-402.
[52] Chou, K. C. and Cai, Y. D. (**2003**) *J. Cellular Biochemistry* (Addendum, ibid. 2004, 91, 1085), 90, 1250-1260.
[53] Chou, K. C. and Cai, Y. D. (**2005**) *Bioinformatics*, *21*, 944-950.
[54] Chou, K. C. and Elrod, D. W. (**1999**) *Proteins: Structure, Function, and Genetics*, *34*, 137-153.
[55] Cai, Y. D., Zhou, G. P., and Chou, K. C. (**2003**) *Biophysical J.*, *84*, 3257-3263.
[56] Chou, K. C. and Cai, Y. D. (**2005**) *J. Chemical Information and Modeling*, *45*, 407-413.
[57] Chou, K. C. and Cai, Y. D. (**2005**), *Biochem. Biophys. Res. Comm.*, *327*, 845-847.
[58] Chou, K. C. and Cai, Y. D. (**2004**) *Prot. Science*, *13*, 2857-2863.
[59] Chou, K. C. (**2005**) *Bioinformatics*, *21*, 10-19.
[60] Chou, K. C. and Elrod, D. W. (**2003**) *J. Proteome Research*, *2*, 183-190.
[61] Chou, K. C. and Cai, Y. D. (**2004**) *Biochem. Biophys. Res. Commun.*, *325*, 506-509.
[62] Chou, K. C. and Cai, Y. D. (**2004**) *Proteins: Structure, Function, and Genetics*, *55*, 77-82.
[63] Cai, Y. D., Zhou, G. P., Jen, C. H., Lin, S. L., and Chou, K. C. (**2004**) *J. Theoretical Biology*, *228*, 551-557.
[64] Elrod, D. W. and Chou, K. C. (**2002**) *Prot. Engineering*, *15*, 713-715.
[65] Chou, K. C. and Elrod, D. W. (**2002**) *J. Proteome Research*, *1*, 429-433.
[66] Chou, K. C. (**2005**) *J. Proteome Research*, *4*, 1413-1418.
[67] Chou, K. C. and Cai, Y. D. (**2003**) *Proteins: Structure, Function, and Genetics*, *53*, 282-289.

Received: October 25, 2005 Revised: November 19, 2005 Accepted: November 22, 2005