Protein Structural Class Determination Using Support Vector Machines

Zerrin Isik, Berrin Yanikoglu, Ugur Sezerman
zisik@su.sabanciuniv.edu, berrin,ugur@sabanciuniv.edu
http://fens.sabanciuniv.edu

Sabanci University, Tuzla, Istanbul, Turkey 34956

Abstract. Proteins can be classified into four structural classes (all-α, all-β, α/β, $\alpha+\beta$) according to their secondary structure composition. In this paper, we predict the structural class of a protein from its Amino Acid Composition (AAC) using Support Vector Machines (SVM).
A protein can be represented by a 20 dimensional vector according to its AAC. In addition to the AAC, we have used another feature set, called the Trio Amino Acid Composition (Trio AAC) which takes into account the amino acid neighborhood information. We have tried both of these features, the AAC and the Trio AAC, in each case using a SVM as the classification tool, in predicting the structural class of a protein. According to the Jackknife test results, Trio AAC feature set shows better classification performance than the AAC feature.

1 Introduction

Protein folding is the problem of finding the 3D structure of a protein, also called its native state, from its amino acid sequence. There are 20 different types of amino acids (labelled with their initials as: A, C, G, ...) and one can think of a protein as a sequence of amino acids (e.g. AGGCT...). Hence the folding problem is finding how this amino acid chain (1D structure) folds into its native state (3D structure). Protein folding problem is a widely researched area since the 3D structure of a protein offers significant clues about the function of a protein which cannot be found via experimental methods quickly or easily.

In finding the 3D structure of a protein, a useful first step is finding the 2D structure, which is the local shape of its subsequences: a helix (called α-helix) or a strand (called β-strand). A protein is classified into one of four *structural classes*, a term introduced by Levitt and Chothia, according to its secondary structure components: all-α, all-β, α/β, $\alpha+\beta$, [1, 2]. An illustration of two of these (all-α, all-β) is given in Figure 1.

The structural class of a protein has been used in some secondary structure prediction algorithms [3–5]. Once, the structural class of a protein is known, it can be used to reduce the search space of the structure prediction problem: most of the structure alternatives will be eliminated and the structure prediction task will become easier and faster.

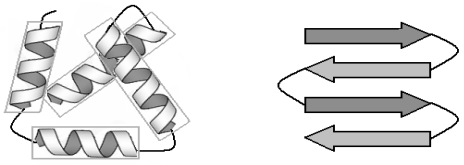

Fig. 1. The illustration of two structural classes. The one on the left is a protein composed of only α-helices whereas the one on the right is composed of what is called a β-sheet (formed by strands of amino acids).

During the past ten years, much research has been done on the structural classification problem [6–18]. Chou [12] used the amino acid composition of a protein and Mahalanobis distance to assign a protein into one of the four structural classes. Due to the high reported performance, Wang et al. tried to duplicate Chou's work using the same data set, without success [22]. More recently, Ding and Dubchak compare the classification performance of ANNs and SVMs on classifying proteins into one of 27 fold classes, which are subclasses of the structural classes [17]. Tan and coworkers also work on the fold classification problem (for 27 fold classes), using a new ensemble learning method [18].

These approaches typically use the *Amino Acid Composition* (AAC) of the protein as the base for classification. The AAC is a 20 dimensional vector specifying the composition percentage for each of the 20 amino acids. Although the AAC largely determines structural class, its capacity is limited, since one looses information by representing a protein with only a 20 dimensional vector. We improved the classification capacity of the AAC by extending it to the *Trio AAC*. The Trio AAC records the occurrence frequency of all possible combinations of consecutive amino acid triplets in the protein. The frequency distribution of neighboring triplets is very sparse because of the high dimensionality of the Trio AAC input vector (20^3). Furthermore, one also should exploit the evolutionary information which shows that certain amino acids can be replaced by the others without disrupting the function of a protein. These replacements generally occur between amino acids which have similar physical and chemical properties [20]. In this work, we have used different clusterings of the amino acids to take into account these similarities and reduce the dimensionality, as explained in Section 2 .

In the results section we compare the classification performance of two feature sets, the AAC and the Trio AAC. The classification performance of a Support Vector Machine with these feature sets is measured on a data set consisting of 117 training and 63 test proteins [12] . The comparison of two different feature sets have proved that the high classification capacity of SVMs and the new feature vector (Trio AAC) lead to much better classification results. Most work in this

area is not directly comparable due to different data sets or different number of classes the proteins are classified into. We use the same data set used by Chou [12] and Wang et al. [22], in order to be able to compare our results to some extent.

2 Protein Structural Class Determination

We have tried two approaches to classify a protein into one of the four structural classes (all-α, all-β, α/β, $\alpha+\beta$). A Support Vector Machine is used with the feature sets of AAC and Trio AAC, which incorporates evolutionary and neighborhood information to the AAC.

We preferred to use a SVM as the classification tool because of its generalization power, as well as its high classification performance on the protein structural classification problem [21, 16, 17]. The SVM is a supervised machine learning technique which seeks an optimal discrimination of two classes, in high dimensional feature space. The superior generalization power, especially for high dimensional data, and fast convergence in training are the main advantages of SVMs. Generally, SVMs are designed for 2-class classification problems whereas our work requires the multi-class classification. Multi-class classification can be achieved using a one-against-one voting scheme, as we have done using the one-against-one voting scheme of the LIBSVM software [23]. In order to get good classification results, the parameters of SVM, especially the kernel type and the error-margin tradeoff (C), should be fixed. In our work, the Gaussian kernels are used since, they provided better separation compared to Polynomial and Sigmoid kernels for all experiments. The value of the parameter C was fixed during the training and later used during the testing. The best performance was obtained with C values ranging from 10 to 100 in various tasks.

We used two different feature sets, the AAC and the Trio AAC, as the input vectors of the SVM. The PDB files were used to form both the AAC and the Trio AAC vectors for the given proteins [24]. After collecting the PDB files of proteins, we extracted the amino acid sequence of each one. The amino acid sequences were then converted to the feature vectors as described in the following sections.

AAC:

The AAC represents protein with a 20 dimensional vector corresponding to the composition (frequency of occurrence) of the 20 amino acids in the protein. Since the frequencies sum up to 1, resulting in only 19 independent dimensions, the AAC can be used as a 19 dimensional vector.

$$X = \left[x_1 \; x_2 \; \ldots \; x_{20} \right] \tag{1}$$

where x_k is the occurrence frequency of the kth amino acid.

Trio AAC:

The Trio AAC is the occurrence frequency of all possible consecutive triplets of amino acids in the protein. Whereas the AAC is a 20-dimensional vector, the Trio AAC vector, consisting of the neighborhood composition of triplets of amino acids, requires a 20x20x20 dimensional vector (e.g. AAA, AAC, ...).

We reduce the dimensionality of the Trio AAC input vector using various different clusterings of the amino acids, also taking into account the evolutionary information. The amino acid clusters are constructed according to hydrophobicity and charge information of amino acids given by Thomas and Dill [20]. We experimented with different number of clusters: 5, 9, or 14 clusters of the amino acids, giving Trio AAC vectors of 125 (5^3), 729 (9^3), and 2744 (14^3) dimensions, respectively.

3 Results

We have measured the performance of two algorithms: SVM with the AAC and SVM with the Trio AAC. We have also compared our test results to another structural classification work which also applied the AAC feature set on the same data set [22]. In all these tests, we have used a data set consisting of 117 training proteins (29-α, 30-β, 29-α/β, 29-$\alpha+\beta$) and 63 (8-α, 22-β, 9-α/β, 24-$\alpha+\beta$) test proteins [12].

A protein is said to belong to a structural class based on the percentage of its α-helix and β-sheet residues. In our data set, the data is labelled according to the following percentage thresholds:

- α class proteins include more than 40% α-helix and less than 5% β-sheet residues
- β class proteins include less than 5% α-helix and more than 40% β-sheet residues
- α/β class proteins include more than 15% α-helix, more than 15% β-sheet, and more than 60% parallel β-sheets
- $\alpha+\beta$ class proteins include more than 15% α-helix, more than 15% β-sheet, and more than 60% antiparallel β-sheets.

Note that the remaining, less-structured parts of a protein, such as loops, are not accounted in the above percentages.

3.1 Training Performance

The term *training performance* is used to denote the performance of the classifer on the training set. Specifically, the training performance is the percentage of the correctly classified training data, once the training completes, and is an indication of how well the training data is learned. Even though what is important is the generalization of a classifier, training performances are often reported for this problem, and we do the same for completeness.

The SVM achieved a near 99.1% training performance for for both sets of features (96.% for β, 100% for the rest). Not achieving a 100% separation on the training data is quite normal and just indicates that the data points may not be linearly separable in the feature space, due to the input space mapping done by the kernel function.

3.2 Test Performance

Table 1 summarizes the test performance of the classifier on the test set (63 proteins), after being trained on the training set (117 other proteins). The AAC and the Trio AAC are used as feature vectors for the SVM.

The average test performances of the SVM using the AAC and the Trio AAC are 71.4% and 66.6%, respectively. The performance of the SVM with Trio AAC feature was found to be lower compared to the AAC feature. This is likely to be due to the high dimensionality of the input data, compared to the size of the training set: if there are points in the test set which are not represented in the training set, they could be misclassified. In this and all the other tables, we report the performance of the Trio AAC using 9 clusters, as that gave the best results.

Table 1. Performance of the classifier on the test set. The AAC feature and the Trio AAC (9 clusters) are used for the SVM.

Class Name	SVM^{AAC}	$SVM^{TrioAAC}$
all-α	100%	100%
all-α	62.5%	62.5%
all-β	77.2%	77.2%
α/β	100%	77.7%
$\alpha+\beta$	58.3%	54.1%
Average	71.4%	66.6%

3.3 Test Performance using the Jackknife Method

The *Jackknife test*, also called the leave-one-out test, is a cross-validation technique which is used when there is a small data set. In the Jackknife test, training is done using all of the data (train + test) leaving one sample out each time; then the performance is tested using that one sample, on that round of train-test cycle. At the end, the test performance is calculated as the average of the test results obtained in all the cycles. This method uses all of the data for testing, but since the test data is not used for the corresponding training phase, the testing is unbiased.

Table 2 displays the results of a Jackknife experiment using both the train and test sets (117 + 63), in conjunction with the AAC and the Trio AAC. According to this Jackknife test results, the performance of the SVM is quite successful. The average classification rates are 85% and 92.7% for the AAC and the Trio AAC, respectively. We achieved the 92.7% classification rate using the Trio AAC which is constructed using 9 amino acid clusters.

Table 2. Jackknife test performance on (117+63) proteins, using the SVM with the AAC and the Trio AAC (9 clusters) features.

	SVM^{AAC}		$SVM^{TrioAAC}$	
Class Name	%	#	%	#
all-α	72.9	(27/37)	72.9	(27/37)
all-β	100	(52/52)	98	(51/52)
α/β	84.2	(32/38)	94.7	(36/38)
$\alpha+\beta$	79.2	(42/53)	100	(53/53)
Average	85.0	(153/180)	92.7	(167/180)

A second Jackknife test has been performed on only the 117 training proteins in order to compare our results to the previous work of Wang and Yuan [22], who also used the AAC feature as a base classifier. The results for both works are shown in Table 3. According to these results, the average classification performance of the SVM (using the AAC) is significantly better than the other work. The average classification rate of the Trio AAC (84.6%) is even better than that of the AAC (74.3%).

Table 3. Jackknife test performance on 117 proteins (the training set only). This experiment was done to compare our results to a previous work of Wang and Yuan (given on the first column), who also used the AAC feature in the Jackknife test on the same proteins [22]. Our results, obtained by the SVM method using the AAC or the Trio AAC, are given on the second and third columns.

Class Name	*Wang et.al.*	SVM^{AAC}	$SVM^{TrioAAC}$
all-α	66.7%	75.8%	82.7%
all-β	56.7%	93.3%	93.3%
α/β	43.3%	71.4%	89.2%
$\alpha+\beta$	46.7%	55.1%	72.4%
Average	53.3%	74.3%	84.6%

4 Summary and Discussion

Despite years of research and the wide variety of approaches that have been utilized, the protein folding problem still remains an open problem. Today the problem is approached in many different directions and divided up into smaller tasks, such as secondary structure prediction, structural class assignment, contact map prediction etc.

In this study, we addressed the structural classification problem and compared the performance of Support Vector Machines using the AAC and the Trio AAC features. The comparison of two feature sets shows that the Trio AAC provides 8-10% improvement in classification accuracy (see Table 2 and 3). We experimented with different number of clusters, 5, 9, and 14 clusters of the amino acids, giving Trio AAC vectors of increasing lengths. The experiment with 9 clusters of the amino acids has the highest classification performance. The better performance of the Trio AAC proves our assumption: the neighborhood and evolutionary information positively contributes on the classification accuracy. We have also obtained better classification rates using more training data, which is as expected.

In literature, there are two studies which use feature vectors similar to the Trio AAC on different domains; however they are on remote homology detection problem and amino acid neighboring effect [25, 26]. We recently became aware of two other studies: Markowetz et al. uses feature vectors similar to the Trio ACC, however the idea of using amino acid clusters (to reduce dimensionality) has not been applied [19]. In this work, 268 protein sequences are classified into a set of 42 structural classes with a 78% performance in cross-validation tests. Cai et al. uses a Support Vector Machine as the classification method and the amino acid composition as feature set and report an average classification performance of 93%, for a set of 204 proteins [16]. However these results are not directly comparable to ours due to the differences in the number of structural classes or in the data sets.

In summary, we devised a new and more complex feature set (Trio AAC) incorporating neighborhood information in addition to the commonly used amino acid composition information. The higher classification rates indicate that the combination of a powerful tool and this new feature set improves the accuracy of the structural class determination problem.

References

1. Levitt, M., Chothia, C.: Structural patterns in globular proteins. Nature **261** (1976) 552–558
2. Richardson, J.S., Richardson, D.C.: Principles and patterns of protein conformation. In Fasman, G.D., ed.: Prediction of protein structure and the principles of protein conformation, New York, Plenum Press (1989) 1–98
3. Deleage, G., Dixon, J.: Use of class prediction to improve protein secondary structure prediction. In Fasman, G.D., ed.: Prediction of protein structure and the principles of protein conformation, New York, Plenum Press (1989) 587–597

4. Kneller, D.G., Cohen, F.E., Langridge, R.: Improvements in protein secondary structure prediction by an enhanced neural network. J Mol Biol **214** (1990) 171–182
5. Eisenhaber, F., Persson, B., Argos, P.: Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. Crit Rev Biochem Mol Biol **30** (1995) 1–94
6. Nakashima, H., Nishikawa, K., Ooi, T.: The folding type of a protein is relevant to the amino acid composition. J Biochem (Tokyo) **99** (1986) 153–162
7. Klein, P., Delisi, C.: Prediction of protein structural class from the amino acid sequence. Biopolymers **25** (1986) 1659–1672
8. Chou, P.Y.: Prediction of protein structural classes from amino acid composition. In Fasman, G.D., ed.: Prediction of protein structure and the principles of protein conformation, New York, Plenum Press (1989) 549–586
9. Zhang, C.T., Chou, K.C.: An optimization approach to predicting protein structural class from amino acid composition. Protein Sci **1** (1992) 401–408
10. Metfessel, B.A., Saurugger, P.N., Connelly, D.P., Rich, S.S.: Cross-validation of protein structural class prediction using statistical clustering and neural networks. Protein Sci **2** (1993) 1171–1182
11. Chandonia, J.M., Karplus, M.: Neural networks for secondary structure and structural class predictions. Protein Sci **4** (1995) 275–285
12. Chou, K.C.: A novel approach to predicting protein structural classes in a (20-1)-d amino acid composition space. Proteins **21** (1995) 319–344
13. Bahar, I., Atilgan, A.R., Jernigan, R.L., Erman, B.: Understanding the recognition of protein structural classes by amino acid composition. Proteins **29** (1997) 172–185
14. Chou, K.C.: A key driving force in determination of protein structural classes. Biochem Biophys Res Commun **264** (1999) 216–224
15. Cai, Y., Zhou, G.: Prediction of protein structural classes by neural network. Biochimie **82** (2000) 783–787
16. Cai, Y.D., Liu, X.J., Xu, X., Chou, K.C.: Prediction of protein structural classes by support vector machines. Comput Chem **26** (2002) 293–296
17. Ding, C.H., Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics **17** (2001) 349–358
18. Tan, A.C., Gilbert, D., Deville, Y.: Multi-class protein fold classification using a new ensemble machine learning approach. Genome Informatics **14** (2003) 206–217
19. Markowetz, F., Edler, L., Vingron, M.: Support vector machines for protein fold class prediction. Biometrical Journal **45** (2003) 377–389
20. Thomas, P.D., Dill, K.A.: An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci U S A **93** (1996) 11628–11633
21. Vapnik, V.: Statistical Learning Theory. NY: Wiley, New York (1998)
22. Wang, Z.X., Yuan, Z.: How good is prediction of protein structural class by the component-coupled method. Proteins **38** (2000) 165–175
23. Chang, C.C., Lin, C.J.: LIBSVM: a Library for Support Vector Machines. (2002)
24. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. Nucleic Acids Res **28** (2000) 235–242
25. Leslie, C., Eskin, E., Noble, W.S.: The spectrum kernel: A string kernel for svm protein classification. In: Pacific Symposium on Biocomputing, Hawaii, USA. (2002)
26. Vishwanathan, S.V.N., Smola, A.J.: Fast kernels for string and tree matching. In: Neural Information Processing Systems: Natural and Synthetic, Vancouver, Canada. (2002)