Classifying G-protein Coupled Receptors with Support Vector Machine*

Ying Huang[1,2,3] and Yanda Li[1,2,3]

[1] Institute of Bioinformatics
[2] MOE Key Laboratory of Bioinformatics
[3] Department of Automation, Tsinghua University Beijing, 10084, China
hying99@mails.tsinghua.edu.cn

Abstract. G-protein coupled receptors (GPCRs) are a class of pharmacologically relevant transmembrane proteins with specific characteristics. They play a key role in different biological process and are very important for understanding human diseases. However, ligand specificity of many receptors remains unknown and only one crystal structure solved to date. It is highly desirable to predict receptor's type using only sequence information. In this paper, Support Vector Machine is introduced to predict receptor's type based on its amino acid composition. The prediction is performed to the amine-binding classes of the rhodopsin-like family. The overall predictive accuracy about 94% has been achieved in a ten-fold cross-validation.

1 Introduction

G-protein-coupled receptors (GPCRs) are a class of pharmacologically relevant proteins characterized by seven transmembrane (7TM) helices. Through their extracellular and transmembrane domains, GPCRs play a key role in a cellular signaling network that regulates many basic physiological processes: neurotransmission, cellular metabolism, secretion, cellular differentiation and growth, inflammatory and immune responses, smell, taste and vision [1]. According to their binding with different ligand types, GPCRs are further classified into different families. Many efforts in pharmaceutical research are current aimed at understanding their structure and function. Despite their importance, there is still only crystal structure solved to date [2]. And many known human GPCRs remain orphans (the activating ligand is unknown) [3]. In contrast, the sequences of thousands of GPCRs are known. So computational methods based on sequence information may be helpful to identify receptor's type [4]. Recently, Elrod and Chou [5] proposed a covariant discriminant algorithm to predict a GPCR's subfamily according to its amino acid composition. In the current study, we try

* This work was funded by the National Natural Science Grant in China (No.60171038 and No.60234020) and the National Basic Research Priorities Program of the Ministry of Science and Technology (No.2001CCA0). Y.H. also thanks Tsinghua University Ph.D. Grant for the support.

F. Yin, J. Wang, and C. Guo (Eds.): ISNN 2004, LNCS 3174, pp. 448–452, 2004.

to apply Support Vector Machine (SVM) method to approach this problem. To demonstrate our algorithm, we apply it to the amine-binding classes of the rhodopsin-like family of GPCRs. There are many medically and pharmacologically important proteins in this family. The results show that the prediction accuracy is significantly improved with this method.

2 Materials and Methods

2.1 Sequence Data

We used the same dataset as that of Erlod and Chou [5], which was taken from GPCRDB [6] (December 2000 release). There are 167 rhodopsin-like amine G-protein-coupled receptors classified into four groups, acetylcholine, adrenoceptor, dopamine and serotonin, as shown in Table 1. Other sub-families of rhodopsin-like amine GPCR are not included, for there are too few sequences in these sub-families to have any statistical significance. Redundancy was reduced so that pair-wise sequence identity is relative low. Accession numbers of these proteins in SWISSPROT can be obtained from Erlod and Chou [5].

Table 1. Summary of 167 receptors classified into four types

Group	Number of Sequences	Average Length
Acetylcholine	31	530.6
Adrenoceptor	44	448.4
Dopamine	38	441.4
Serotonin	54	433.7
Overall	167	457.3

2.2 Data Representation

Protein sequences are strings make up of 20 different amino acids (alphabets). To apply machine learning method such as SVM, we have to extract a fixed length feature vector from protein sequence with variable length. Following Chou [7], a protein is represented by its amino acid composition, corresponding to a vector in the 20-D (dimensional) space.

$$\mathbf{X_k} = \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,20} \end{bmatrix} \qquad (k = 1, 2, \ldots, 167) \tag{1}$$

where $x_{k,1}, x_{k,2}, \ldots, x_{k,20}$ are the 20 components of amino acid composition for the kth protein X_k.

2.3 Support Vector Machine

SVM is a popular machine learning algorithm based on recent advances in statistical learning theory [8,9]. This algorithm first maps data into a high-dimensional feature space, and then constructs a hyperplane as the decision surface between positive and negative patterns. The actual mapping is achieved through a kernel function, making it easy to implement and fast to compute. Several popular kernel functions are:

$$\text{linear kernel: } K(x_i, x_j) = x_i^T x_j \tag{2}$$

$$\text{polynomial kernel: } K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0 \tag{3}$$

$$\text{RBF kernel: } K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2), \gamma > 0 \tag{4}$$

In principle, SVM is a two-class classifier. With the recent improvements, the SVM can directly cope with multi-class classification problem now [10]. The software used to implement SVM was libSVM [11], which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm. LibSVM used "one-against-one" approach to deal with multi-class problem, in which $k(k-1)/2$ classifiers are constructed and each one trains data from two different classes.

2.4 Prediction Accuracy Assessment

The prediction performance was examined by the ten-fold cross-validation test, in which the dataset of 167 GPCRs for the four groups were divided into ten subsets of approximately equal size. We train our algorithm based on nine of these subsets, and then the remaining subset is used to estimate the predictive error of the trained classifier. This process is repeated ten times so that every subset is once used as the test data. The average of the ten estimates of the predictive error rate is reported as the cross-validation error rate. The prediction quality is then evaluated by the overall prediction accuracy and prediction accuracy for each group.

$$\text{overall accuracy} = \sum_{s=1}^{k} p(s)/N \tag{5}$$

$$\text{accuracy(s)} = p(s)/obs(s) \tag{6}$$

Where N is the total number of proteins in the data set (N=167), k is the number of groups (k=4), obs(s) is the number of sequences observed in group s, and p(s) is the number of correctly predicted sequences in group s.

3 Results

3.1 SVM Kernel and Parameters Selection

Experiments have been done on three popular kernel functions: the linear kernel, the polynomial kernel and the RBF kernel. We found that RBF kernel performed

better than linear kernel and polynomial kernel. For RBF kernel, there are two parameters to be selected: kernel parameter γ and penalty parameter C. We choose $\gamma = 0.05$, which is the default value of libSVM software. As for C, various values ranging from 1 to 64 have been tested. The best result was achieved when $C = 4$.

3.2 Comparison with Existing Methods

The SVM prediction performance was compared covariant discriminant algorithm that also based on amino acid compositions. The results are summarized in Table 2. The overall success rate of SVM algorithm is 94.01%, which is about 10% higher than that of the covariant discriminant algorithm. We also observed that SVM improve the prediction performance for every group. Especially for acetylcholine type, SVM improve the accuracy significantly.

Table 2. Performance comparison between covariant discrimination and SVM algorithm.

Group	Cov[a] Accuracy(%)	SVM Accuracy(%)
Acetylcholine	67.74	100
Adrenoceptor	88.64	90.91
Dopamine	81.58	94.74
Serotonin	88.89	92.59
Overall	83.23	94.01

[a] covariant discrimination algorithm

4 Conclusion

In this paper, we introduced SVM method for recognizing the family of GPCRs. The rate of correct identification obtained in ten-cross validation is about 94% for four group classification, which is superior to existing algorithms. This result implies that we can predict the type of GPCRs to a considerably accurate extent using amino acid composition. It is anticipated that our method would be a useful tool for classification of orphan GPCRs and facilitate drug discovery for psychiatric and schizophrenic diseases.

References

1. Hebert, T., Bouvier, M.: Structural and functional aspects of g protein-coupled receptor oligomerization. Biochem Cell Biol **76** (1998) 1–11
2. Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le Trong, I., Teller, D.C., Okada, T., Stenkamp, R.E., Yamamoto, M., Miyano, M.: Crystal structure of rhodopsin: A g protein-coupled receptor. Science **289** (2000) 739–45

3. Schoneberg, T., Schulz, A., Gudermann, T.: The structural basis of g-protein-coupled receptor function and dysfunction in human diseases. Rev Physiol Biochem Pharmacol **144** (2002) 143–227
4. Gaulton, A., Attwood, T.K.: Bioinformatics approaches for the classification of g-protein-coupled receptors. Curr Opin Pharmacol **3** (2003) 114–20
5. Elrod, D.W., Chou, K.C.: A study on the correlation of g-protein-coupled receptor types with amino acid composition. Protein Engineering **15** (2002) 713–715
6. Horn, F., Weare, J., Beukers, M.W., Horsch, S., Bairoch, A., Chen, W., Edvardsen, O., Campagne, F., Vriend, G.: Gpcrdb: an information system for g protein-coupled receptors. Nucleic Acids Res **26** (1998) 275–9
7. Chou, K.C.: A novel approach to predicting protein structural classes in a (20-1)-d amino acid composition space. Proteins Struct. Funct. Genet. **21** (1995) 319–344
8. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, N.Y. (1995)
9. Vapnik, V.N.: Statistical Learning Theory. Wiley, New-York (1998)
10. Hsu, C.W., Lin, C.J.: A comparison of methods for multi-class support vector machines. IEEE Trans. Neural Networks **13** (2002) 415–25
11. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.