Predicting the Coupling Specificity of G-protein Coupled Receptors to G-proteins by Support Vector Machines

Cui-Ping Guan, Zhen-Ran Jiang, and Yan-Hong Zhou*

Hubei Bioinformatics and Molecular Imaging Key Laboratory, Huazhong University of Science and Technology, Wuhan 430074, China.

G-protein coupled receptors (GPCRs) represent one of the most important classes of drug targets for pharmaceutical industry and play important roles in cellular signal transduction. Predicting the coupling specificity of GPCRs to G-proteins is vital for further understanding the mechanism of signal transduction and the function of the receptors within a cell, which can provide new clues for pharmaceutical research and development. In this study, the features of amino acid compositions and physiochemical properties of the full-length GPCR sequences have been analyzed and extracted. Based on these features, classifiers have been developed to predict the coupling specificity of GPCRs to G-proteins using support vector machines. The testing results show that this method could obtain better prediction accuracy.

Key words: GPCR, G-protein, SVM, coupling specificity

Introduction

G-protein coupled receptors (GPCRs) represent one of the most important classes of drug targets for pharmaceutical industry and play important roles in cellular signal transduction as cell surface receptor proteins (*1*, *2*). A GPCR only has a single polypeptide that consists of seven transmembrane α-helices, three extracellular and three intracellular loops connecting the transmembrane domains. The N-terminal of the polypeptide is located on the extracellular side and the C-terminal extends to the cytoplasm. GPCRs can be activated by various extracellular signaling molecules that bind to the receptors and trigger their conformational changes. The activated receptors will then couple with G-proteins ($G_{i/o}$, $G_{q/11}$, G_s, and $G_{12/13}$) and further activate different signaling pathways (*3*). Since most GPCRs are coupled with one single subtype of G-proteins, they are broadly categorized into $G_{i/o}$-, $G_{q/11}$-, G_s-, and $G_{12/13}$-coupled receptors based on their G-protein coupling preference.

Predicting the coupling specificity of GPCRs to G-proteins is vital for further understanding the mechanism of signal transduction and the function of the receptors within a cell, which may provide new clues for pharmaceutical research and development. Many computational methods such as the pattern discovery-based method (*4*), the naive Bayes model (*5*), the chemometric approach (*6*), and the hidden Markov model (HMM) (*7–9*), have been developed. Most of these methods are only based on the sequence information of GPCRs. Furthermore, since previous experimental researches have demonstrated that the coupling specificity of GPCRs to G-proteins is mainly determined by the intracellular domains of the receptor (*10*, *11*), the reported computational methods only used the intracellular regions of GPCR sequences as the information for the development of the prediction models.

In this study, we hypothesize that, besides the intracellular regions of GPCRs, other factors such as the physiochemical properties of amino acid residues, the extracellular and transmembrane regions of GPCRs, might also have some influence on the coupling specificity, and therefore can be used as additional information to further improve the prediction accuracy. Accordingly, the features of amino acid compositions and the physiochemical properties of the full-length GPCR sequences have been analyzed and extracted, and classifiers have been developed to predict the coupling specificity based on these features and support vector machines (SVMs). The testing results show that this method could obtain better prediction accuracy.

*** Corresponding author.**
E-mail: yhzhou@hust.edu.cn

Results

In this study, 124 human GPCR sequences (62, 33, and 29 sequences for $G_{i/o}$-, $G_{q/11}$-, G_s-class, respectively) were collected from the gpDB database (*12*), which were used to develop the SVM classifiers for three classes of GPCRs. Ten-fold cross-validation process was used to evaluate the performance of the proposed method, that is, nine tenths of the sequences were used as the training set to extract classification features and train SVM classifiers, and one tenth were used as the test set to verify the prediction results. The training and test sets were segregated at random and the process was repeated 30 times. Three measures, sensitivity (Sn), specificity (Sp), and overall performance accuracy (Acc), were used to evaluate the prediction performance. Sn and Sp are defined as Sn = TP/(TP+FN) and Sp = TN/(TN+FP), respectively, where TP is the number of correctly predicted positive samples for a specific class of GPCRs, FN is the number of incorrectly predicted positive samples, and FP is the number of incorrectly predicted negative samples. The overall performance accuracy, which measures the average accuracy of predicting the three classes of GPCRs, is defined as the percentage of all correctly predicted number of positive samples to the total number of positive samples.

During the development of SVM classifiers, different kernel functions were tried, and it was found that using the linear kernel function could get the best results. Besides the commonly used feature of amino acid compositions, additional features such as hydrophobicity, normalized van der Waals volume, polarity, polarizability, and the charge of amino acids were also used in this study for better prediction results. For each of these additional features, it is defined as the feature frequency in the form of k-mers (see Materials and Methods) and the value of parameter k ($k \geq 1$) has a great influence on the prediction results. The results of predicting the coupling specificity for the three classes of GPCRs with different k (k= 1, 2, and 3) are given in Table 1, which were obtained by using the linear kernel function and the features extracted from the full-length GPCR sequences.

It can be seen that, when k= 2, the overall performance accuracy is higher than others, while the Sn and Sp for all three classes also achieve a relatively high level. In order to compare the prediction performance based on the information of the full-length sequences of GPCRs with that only based on the intracellular part as adopted by other researches, we predicted the transmembrane helices of these GPCRs with the programs ConPred II (*13*) and HMMTOP 2.0 (*14*), then extracted the intracellular sequences and converted them into fixed-length feature vectors with the same method. The prediction results with k= 2 and k= 3 are given in Table 2.

Discussion

The primary aim of this study is to improve the accuracy of predicting the coupling specificity of GPCRs to G-proteins by such measures as making use of the

Table 1 Predicting Results Based on Features of the Full-Length Sequences

Class	k=1		k=2		k=3	
	Sn (%)	Sp (%)	Sn (%)	Sp (%)	Sn (%)	Sp (%)
$G_{i/o}$	91.67	90.11	95.42	94.61	95.56	91.15
$G_{q/11}$	80.00	80.83	83.34	87.54	79.89	93.89
G_s	65.00	83.33	81.25	90.00	73.67	92.78
Acc (%)	82.34		88.89		86.60	

Table 2 Predicting Results Based on Features of the Intracellular Sequences

Class	k=2		k=3	
	Sn (%)	Sp (%)	Sn (%)	Sp (%)
$G_{i/o}$	89.17	91.01	90.00	87.06
$G_{q/11}$	73.34	81.00	70.00	87.54
G_s	67.50	79.61	65.00	65.00
Acc (%)	79.91		78.83	

information of full-length GPCR sequences, integrating the compositional features and physiochemical properties of amino acids, and using the SVM method. By comparing Table 1 with Table 2, it is distinct that the prediction accuracy based on the full-length sequences of GPCRs is better than that only based on the intracellular part, meaning that the full-length sequences do contain more information about the coupling specificity of GPCRs to G-proteins than the intracellular sequences. These results justify our hypothesis. In addition, the methods based on the intracellular sequences need to extract them from the full-length sequences of GPCRs using transmembrance topology prediction programs, which may bring some errors as the accuracy of predicting the seven transmembrance α-helices of GPCRs is currently only about 80%. On the contrary, this problem does not exist in the proposed method that uses the information of full-length GPCR sequences directly. Furthermore, the additional features used in this study, including hydrophobicity, normalized van der Waals volume, polarity, polarizability, and the charge of amino acids, can contribute to better prediction accuracy (data not shown), implying that the physiochemical properties of amino acids might also play important roles in determining the coupling specificity of GPCRs to G-proteins.

It is worth noticing that the overall prediction accuracy of our method is still affected by the following factors. Firstly, the size of dataset and the imbalance among different GPCR classes have a great influence on the prediction accuracy. Secondly, although the collected dataset has been filtrated to only include single coupled GPCRs, it may still contain some potential promiscuous receptors, which could couple to more than one class of G-proteins and have not been validated by biological experiments. On one hand, if the training set contains the promiscuous receptors, it will influence the feature extraction of the single coupled receptors, therefore, the test set cannot be classified effectively. On the other hand, if the test set has promiscuous ones, it will not obtain the overall prediction results of their coupling specificity to G-proteins. Fortunately, with the accumulation of more experimental data, the influence of the above factors on the prediction accuracy will decrease gradually.

In a word, the testing results demonstrate that the method proposed in this study could obtain better prediction accuracy. Future work of this study will focus on exploring and integrating more features of GPCRs to further improve the accuracy of predicting the coupling specificity of GPCRs to G-proteins and to develop novel approaches to distinguish single coupled receptors from promiscuous ones.

Materials and Methods

Dataset

A set of human GPCR sequences with known coupling specificity was collected from the gpDB database (http://bioinformatics.biol.uoa.gr/gpDB). This database, which is useful for the study of GPCR/G-protein interactions, contains 469 species of G-proteins and GPCR sequences. We selected 124 human GPCRs that meet the following criteria: Firstly, only single coupled receptors (only couple to one class of G-proteins) were included in the dataset. $G_{12/13}$-coupled receptors were not included because of insufficient data. Secondly, those GPCR sequences labeled with "fragment" were excluded. Finally, the GPCRs were divided into three groups according to their coupling specificity, and the $G_{i/o}$-, $G_{q/11}$-, and G_s-coupled receptors contained 62, 33, and 29 sequences, respectively. Furthermore, all of these sequences were verified with the TiPS (*15*) and SWISS-PROT database.

Generally, in the process of constructing the positive and negative samples for developing classifiers, two kinds of methods are wildly used, one is the one-against-other (1-v-n) method, and the other is the one-against-one (1-v-1) method. Supposing there are N classes to be classified, the one-against-other method means to pick up the samples in one class as positive and the ones in remnant classes as negative. Another method, one-against-one, uses the samples in one class as positive and the ones in another class as negative. In this study, the numbers of GPCR sequences for the classes $G_{q/11}$ and G_s are not large enough, which may cause the imbalance problem in the dataset. The one-against-other method will further aggravate the imbalance problem, while the one-against-one method can improve the prediction accuracy for classes with fewer samples, and therefore was adopted in this study.

Coding of GPCR sequences

For each GPCR sequence, the feature vector was assembled from the encoded representations of amino acid compositions, hydrophobicity, normalized van der Waals volume, polarity, polarizability, and charge.

Two different coding schemes were adopted for the amino acid compositions and the five physiochemical properties, respectively.

According to the amino acid compositions, a protein sequence is represented by a vector in a 20-dimensional space:

$$\overrightarrow{x}_a = [f_1, f_2, \ldots, f_{20}]^{\mathrm{T}} \tag{1}$$

where f_i $(i = 1, 2, \ldots, 20)$ is the occurrence frequency of the twenty amino acids in the sequence.

For each of the five physiochemical properties, the twenty amino acids are grouped into three classes represented by P_i $(i = 1, 2,$ and $3)$ according to their different values. Each amino acid is replaced by P_i. So a primal sequence is transformed into five different physiochemical sequences consisting of P_1, P_2, and P_3, which are called $PCseq$. Then, the k-mer vector C and the distribution vector D for the $PCseq$ corresponding to each of the five physiochemical properties are calculated, respectively (*16*, *17*).

Given an integer $k \geq 1$, k-mers are composed of all the continuous subsequences with length k. There are possibly 3^k permutation and combination of the subsequences. So the k-mer vector C is a vector with 3^k dimensions and the value of each dimension is the occurrence frequency of every subsequence in the $PCseq$. The distribution vector D, which is used to describe the global distribution of the property in the $PCseq$, is described by five chain lengths (in percent), within which 25%, 50%, 75%, and 100% of the amino acids with a certain class P_i in each of the five properties are contained. It is defined as:

$$D = \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{Pos\big(L(i) \times j \times 25\%\big)}{L(seq)} \times 100\% \tag{2}$$

where $L(i)$ is the number of amino acids in the special class P_i, and $L(seq)$ is the length of $PCseq$. For each of the chosen physiochemical properties, its k-mer vector C and the distribution vector D were calculated and combined.

Finally, a protein sequence was converted into a vector with $[20 + (3^k + 4 \times 3) \times 5)]$ dimensions as the input of the SVM classifiers.

SVM

SVM is a standard supervised learning algorithm based on recent developments in the statistical learning theory (*18*, *19*). It builds a hyperplane separating the positive and negative examples in multiple-dimensional space. The SVM calculation was implemented by using the LIBSVM 2.8 (*20*) software package. The software enables the user to choose different parameters and kernel functions including linear kernel function, radial basis function, and polynomial kernel function to obtain the best effect. In this study, three SVMs were constructed for classifying the $\mathrm{G_{i/o}}$-, $\mathrm{G_{q/11}}$-, and $\mathrm{G_s}$-GPCRs. The comparison results of using different kernel functions show that the linear kernel function can achieve the best accuracy, which is therefore used in the developed SVM classifiers.

In addition, the penalty factor C is an important parameter of SVM, which has a great influence on the prediction results. In this study, the optimal value of this parameter was searched in the range of 1 to 200 by comparing the prediction results.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 90203011 and 30370354) and the Ministry of Education of China (No. 505010 and CG2003-GA002).

References

1. LeVine, H. 3rd. 1999. Structural features of heterotrimeric G-protein-coupled receptors and their modulatory proteins. *Mol. Neurobiol.* 19: 111-149.
2. Elrod, D.W. and Chou, K.C. 2002. A study on the correlation of G-protein-coupled receptor types with amino acid composition. *Protein Eng.* 15: 713-715.
3. Bockaert, J. and Pin, J.P. 1999. Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J.* 18: 1723-1729.
4. Moller, S., *et al.* 2001. Prediction of the coupling specificity of G protein coupled receptors to their G proteins. *Bioinformatics* 17: S174-181.
5. Cao, J., *et al.* 2003. A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. *Bioinformatics* 19: 234-240.
6. Henriksson, Å. 2003. Prediction of G-protein coupling of GPCRs—a chemometric approach. Master of Science thesis. Department of Molecular Biology, Linköping University, Linköping, Sweden.
7. Qian, B., *et al.* 2003. Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs. *FEBS Lett.* 554: 95-99.

8. Sreekumar, K.R., *et al.* 2004. Predicting GPCR–G-protein coupling using hidden Markov models. *Bioinformatics* 20: 3490-3499.
9. Sgourakis, N.G., *et al.* 2005. A method for the prediction of GPCRs coupling specificity to G-proteins using refined profile Hidden Markov Models. *BMC Bioinformatics* 6: 104.
10. Wess, J. 1998. Molecular basis of receptor/G-protein-coupling selectivity. *Pharmacol. Ther.* 80: 231-264.
11. Wong, S.K. 2003. G protein selectivity is regulated by multiple intracellular regions of GPCRs. *Neurosignals* 12: 1-12.
12. Elefsinioti, A.L., *et al.* 2004. A database for G proteins and their interaction with GPCRs. *BMC Bioinformatics* 5: 208.
13. Arai, M., *et al.* 2004. ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucleic Acids Res.* 32: W390-393.
14. Tusnady, G.E. and Simon, I. 2001. The HMMTOP transmembrane topology prediction server. *Bioinformatics* 17: 849-850.
15. Alexander, S., *et al.* 1999. TiPS receptor and ion channel nomenclature supplement 1999. *Trends Pharmacol. Sci.* 19: S1.
16. Dubchak, I., *et al.* 1995. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. USA* 92: 8700-8704.
17. Cai, C.Z., *et al.* 2004. Enzyme family classification by support vector machines. *Proteins* 55: 66-76.
18. Leslie, C., *et al.* 2002. The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomput.* pp.564-575.
19. Cristianini, N. and Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, Cambridge, UK.
20. Fan, R.E., *et al.* 2005. Working set selection using the second order information for training SVM. *J. Mach. Learn. Res.* 6: 1889-1918.