Prediction and Classification of Human G-protein Coupled Receptors Based on Support Vector Machines

Yun-Fei Wang, Huan Chen, and Yan-Hong Zhou*

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

A computational system for the prediction and classification of human G-protein coupled receptors (GPCRs) has been developed based on the support vector machine (SVM) method and protein sequence information. The feature vectors used to develop the SVM prediction models consist of statistically significant features selected from single amino acid, dipeptide, and tripeptide compositions of protein sequences. Furthermore, the length distribution difference between GPCRs and non-GPCRs has also been exploited to improve the prediction performance. The testing results with annotated human protein sequences demonstrate that this system can get good performance for both prediction and classification of human GPCRs.

Key words: GPCR, prediction, classification, SVM

Introduction

G-protein coupled receptors (GPCRs) are a class of transmembrane proteins that can be bound by structurally diverse ligands to activate a variety of cellular signaling cascades, which play important roles in cellular signal transduction. For the pharmaceutical industry, GPCRs are one of the most important classes of drug targets. More than 50% of drugs currently available on the market act through GPCRs (*1*). Human GPCRs can be grouped into four families as A, B, C, and frizzled/smoothened (fz_smo) based on the specificity of their ligands (*2*, *3*). Although a number of GPCRs have been identified in the human genome, there is still room for finding novel GPCRs. Furthermore, hundreds of the identified GPCRs still remain orphaned (with unknown ligand specificity) or poorly characterized. Computational methods are frequently used to facilitate the identification and characterization of receptors, which could lead to the discovery of novel signal transduction pathways and provide new insights into the disease process and drug discovery.

A number of strategies have been developed for the prediction and classification of GPCRs, including the pairwise sequence alignment method (*4*), the Bayes network method (*5*), the hidden Markov model (HMM) method (*6*, *7*), and the support vector machine (SVM) method (*1*, *8*, *9*). All of these methods have been widely used for solving various problems related to biological sequences. For the prediction and classification of GPCRs, however, the performance of these methods is still not very effective owing to the complexity that many GPCRs with analogous functions have resulted from convergent evolution (*7*) and there might exist some higher-order relationships between GPCR sequences and their functions (*10*).

In this study, the SVM method and a three-step strategy were applied to develop a computational system for the prediction and classification of human GPCRs. Firstly, the length distribution difference between human GPCRs and non-GPCRs was analyzed and a length-based filtration system was constructed to filter out the sequences whose lengths are quite different from those of GPCRs. Secondly, an SVM-based GPCR prediction system was developed to discriminate GPCRs from non-GPCRs using compositional features of protein sequences. Finally, an SVM-based GPCR classification system was constructed to further judge which subfamily a potential GPCR sequence might belong to.

Results and Discussion

Length distribution of GPCRs

Using all of the annotated human protein sequences (including 653 GPCRs and 10,845 non-GPCRs)

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

provided by UniProt (ftp://cn.expasy.org/databases/uniprot), the length distributions of GPCRs and non-GPCRs were analyzed in this study (Figure 1). It is clear that the length distribution of human GPCRs is quite different from that of non-GPCRs.

Based on the above length distribution difference between GPCRs and non-GPCRs, a filtration method was adopted to filter out the sequences whose lengths are quite different from those of GPCRs. By this method, nearly 1/3 non-GPCRs (3,119 out of 10,845) can be filtered out at the cost of missing only about 0.6% of GPCRs (4 out of 653). Furthermore, some of the filtered non-GPCRs belong to the non-GPCR transmembrane proteins that are usually difficult to be discriminated from GPCRs. Therefore, this length-based filtration method can improve the performance of GPCR prediction.

Performance of GPCR prediction

For the prediction of GPCRs, an SVM-based system was developed using statistically filtered single amino acid and dipeptide composition features of protein sequences. The dataset used to train this system consisted of 597 GPCRs (positive samples) and 1,825 non-GPCRs (negative samples). To evaluate its prediction performance, this system was applied to discriminate the 653 human GPCRs from 10,845 non-GPCRs. Three measures including sensitivity (Sn), specificity (Sp), and Matthews correlation coefficient (MCC) were utilized to evaluate the prediction accuracy. Let TP (true positive) and TN (true negative) be the number of correctly predicted positive and negative samples, FP (false positive) and FN (false negative) be the number of incorrectly predicted positive and negative samples, respectively, then Sn, Sp, and MCC are defined as:

$$Sn = TP/(TP + FN)$$
$$Sp = TN/(TN + FP)$$
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$$

The results of discriminating 653 human GPCRs from 10,845 non-GPCRs are shown in Figure 2.

The results demonstrate that our GPCR prediction system can discriminate human GPCRs from non-GPCRs with a specificity of 97.2%, a sensitivity of 95.4%, and an MCC value of 0.96. This performance is very close to that of the method developed by Bhasin and Raghava (*1*) using dipeptide composition features and evaluated with the same datasets. In addition, further analysis of prediction results reveals that the false positives are mainly other transmembrane proteins, suggesting that how to distinguish GPCRs from other transmembrane proteins is important for further improving the performance of GPCR prediction.

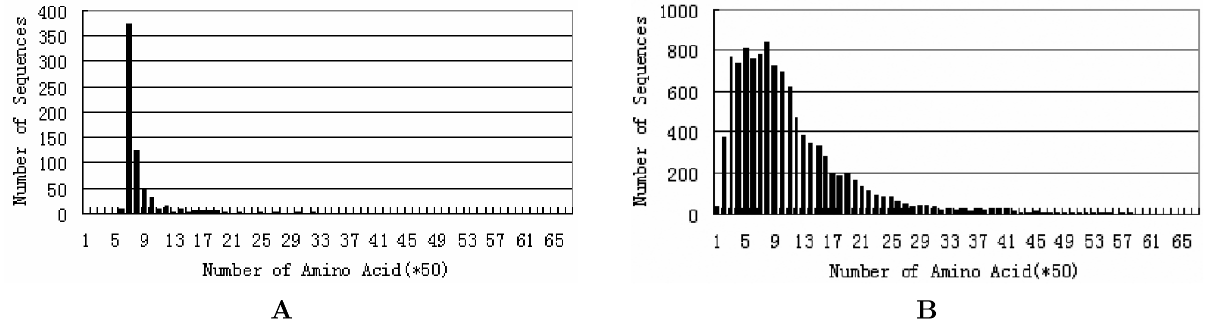

Fig. 1 The length distributions of human GPCRs (**A**) and non-GPCRs (**B**).

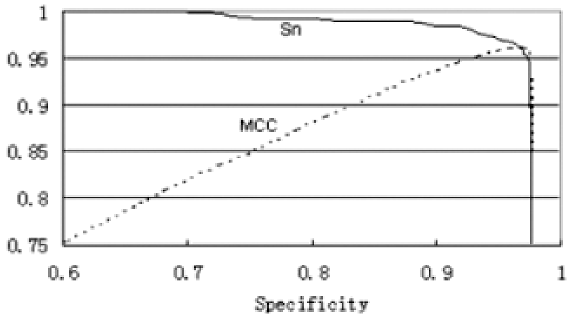

Fig. 2 The results of discriminating 653 human GPCRs from 10,845 non-GPCRs.

Performance of GPCR classification

For the classification of GPCRs, an SVM-based system was developed using statistically filtered single amino acid, dipeptide, and tripeptide composition features of protein sequences. A dataset containing all of the 451 ligand-known huamn GPCRs (395 for family A, 30 for family B, 15 for family C, and 11 for family fz_smo) was used to train and test this system. The measures to evaluate the classification performance include Sn, Sp, and accuracy (Acc), which is defined as:

$$Acc = (TP + TN)/(TP + TN + FP + FN)$$

The classification performance of this system is given in Table 1. For the purpose of comparison, the performance of Bhasin and Raghava's method (*1*) when tested with the same data and evaluated with the same measures is also shown in Table 1. The results demonstrate that the classification performance of our system is better than that of Bhasin and Raghava's method that only exploits dipeptide composition features, suggesting that our method, which selects suitable features from single amino acid, dipeptide, and tripeptide composition features, is more effective for GPCR classification.

Materials and Methods

System flowchart for GPCR prediction and classification

Figure 3 is the system flowchart for GPCR prediction and classification, which includes three steps: (1) Using a length-based filtration system to filter out sequences whose lengths are quite different from those of GPCRs. (2) Using a GPCR prediction system to discriminate GPCRs from non-GPCRs. (3) Using a GPCR classification system to judge which subfamily a potential GPCR sequence might belong to.

Table 1 Performance Comparison of Human GPCR Classification

Family	Our method			Bhasin and Raghava's method		
	Acc	Sp	Sn	Acc	Sp	Sn
A	0.995	0.992	1	0.995	0.992	1
B	0.989	1	0.967	0.966	1	0.900
C	1	1	1	0.897	1	0.400
fz_smo	0.979	1	0.917	0.938	1	0.750

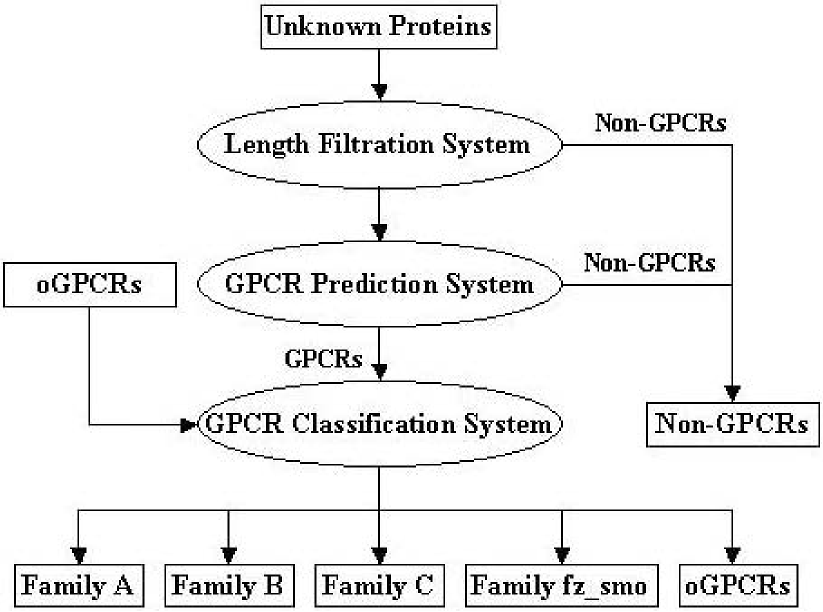

Fig. 3 System flowchart for GPCR prediction and classification.

Length-based filtration system

According to the length distribution difference between human GPCRs and non-GPCRs as shown in Figure 1, a simple determinant function is currently adopted in the length-based filtration system to filter out sequences whose lengths are quite different from those of GPCRs. That is, protein sequences shorter than 250 amino acids or longer than 1,600 amino acids are considered as non-GPCRs. However, it might be a better solution to build a probabilistic model for the length distribution and to consider it as an additional feature in the GPCR prediction system.

GPCR prediction system

The GPCR prediction system is based on an SVM model with the feature vector consisting of statistically filtered single amino acid and dipeptide composition features of protein sequences. The datasets used to develop this system consisted of 597 positives samples selected from all 653 annotated human GPCRs by excluding fragments and redundant sequences, and 1,825 negative samples randomly selected from human non-GPCRs according to the length distribution of GPCRs. The selection of suitable features is the kernel for developing an effective prediction model. Previous studies have revealed that the dipeptide composition feature seems to be significant for discriminating GPCRs from non-GPCRs (*1*). However, we found that some of the 400 dipeptide compositions exhibit insignificant differences between GPCRs and non-GPCRs, and that adding single amino acid composition features could improve the prediction performance of SVM models. To select significant features for GPCR prediction, the following method was used in this study to calculate the significance value for each of the single amino acid and dipeptide composition features. For a feature x, its significance value u is defined as:

$$u(x) = \frac{|\mu(x)_G - \mu(x)_N|}{\sqrt{\frac{\sigma(x)_G^2}{n_G} + \frac{\sigma(x)_N^2}{n_N}}}$$

where G and N denote GPCRs and non-GPCRs, respectively; μ is the average frequency of feature x; σ is the variance of feature x; and n is the number of samples. The u value distributions for single amino acid and dipeptide composition features are shown in Figures 4 and 5, respectively. In this study, features with u value higher than 10 were selected for developing the GPCR prediction model.

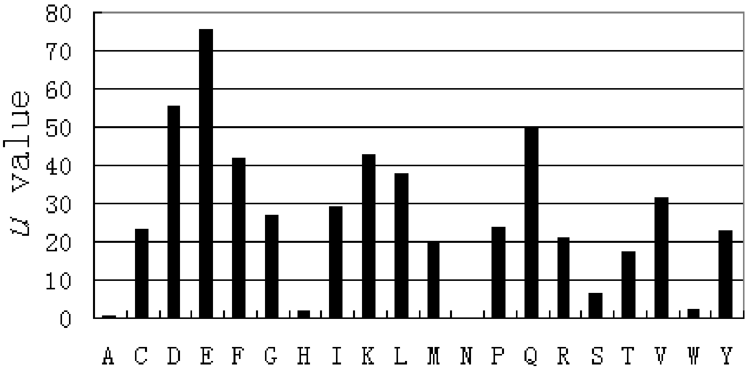

Fig. 4 The μ value distribution of single amino acid composition features.

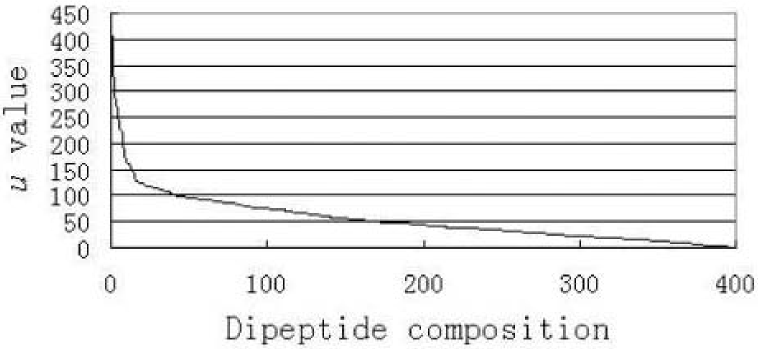

Fig. 5 The μ value distribution of dipeptide composition features.

In this study, the SVM prediction model was developed using the freely available software SVM_light (*11*). The radial basis function (RBF) was chosen as the kernel, and the two related parameters γ and C were set to 0.125 and 3, respectively.

GPCR classification system

For the classification of GPCRs, four "one-versus-rest" SVM models were developed for the four GPCR subfamilies, respectively. A GPCR sequence will be predicted as a member of the subfamily whose SVM model can get the highest output for this sequence. GPCR sequences rejected by all subfamily SVM models will be classified as orphan GPCRs (oGPCRs). To obtain better classification performance, the features used to develop the above SVM models were selected from single amino acid, dipeptide, and tripeptide composition features using the same method as described in the last section.

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. Bhasin, M. and Raghava, G.P. 2004. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res.* 32: W383-389.
2. Horn, F., *et al.* 1998. GPCRDB: an information system for G-protein coupled receptors. *Nucleic Acids Res.* 26: 275-279.
3. Vassilatis, D.K., *et al.* 2003. The G protein-coupled receptor repertoires of human and mouse. *Proc. Natl. Acad. Sci. USA* 100: 4903-4908.
4. Lapinsh, M., *et al.* 2002. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Sci.* 11: 795-805.
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. Möller, S., *et al.* 2001. Prediction of the coupling specificity of G protein coupled receptors to their G proteins. *Bioinformatics* 17: S174-181.
7. Papasaikas, P.K., *et al.* 2003. A novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden Markov models. *SAR QSAR Environ. Res.* 14: 413-420.
8. Karchin, R., *et al.* 2002. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* 18: 147-159.
9. Bhasin, M. and Raghava, G.P. 2005. GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors. *Nucleic Acids Res.* 33: W143-147.
10. Fredriksson, R. and Schioth, H.B. 2005. The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol. Pharmacol.* 67: 1414-1425.
11. Joachims, T. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods: Support Vector Learning* (eds. Schölkopf, B., *et al.*). MIT Press, Cambridge, USA.