Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W383–W389***
DOI: 10.1093/nar/gkh416

GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors

Manoj Bhasin and G. P. S. Raghava*

Institute of Microbial Technology Sector 39-A, Chandigarh, 160036, India

Received February 12, 2004; Revised and Accepted April 2, 2004

ABSTRACT

G-protein coupled receptors (GPCRs) belong to one of the largest superfamilies of membrane proteins and are important targets for drug design. In this study, a support vector machine (SVM)-based method, GPCRpred, has been developed for predicting families and subfamilies of GPCRs from the dipeptide composition of proteins. The dataset used in this study for training and testing was obtained from http://www.soe.ucsc.edu/research/compbio/gpcr/. The method classified GPCRs and non-GPCRs with an accuracy of 99.5% when evaluated using 5-fold cross-validation. The method is further able to predict five major classes or families of GPCRs with an overall Matthew's correlation coefficient (MCC) and accuracy of 0.81 and 97.5% respectively. In recognizing the subfamilies of the rhodopsin-like family, the method achieved an average MCC and accuracy of 0.97 and 97.3% respectively. The method achieved overall accuracy of 91.3% and 96.4% at family and subfamily level respectively when evaluated on an independent/blind dataset of 650 GPCRs. A server for recognition and classification of GPCRs based on multiclass SVMs has been set up at http://www.imtech.res.in/raghava/gpcrpred/. We have also suggested subfamilies for 42 sequences which were previously identified as unclassified ClassA GPCRs. The supplementary information is available at http://www.imtech.res.in/raghava/gpcrpred/info.html.

INTRODUCTION

G-protein coupled receptors (GPCRs) constitute a vast family of cell surface receptor proteins that are central to the signaling network which regulates the basic cellular processes (1). GPCRs consist of a single polypeptide that crosses the membrane seven times (2). The N-terminal of these proteins is located extracellularly and the C-terminal extended in the cytoplasm. This arrangement makes these proteins capable of transducing an extracellular signal into the cell via a guanine binding protein (G-protein) (3). This signal is crucial for the regulation of a large number of metabolic processes such as neurotransmission, hormonal secretion, cellular differentiation and metabolism (4). Therefore, structural and functional annotation of these proteins is useful in understanding the processes of signal transduction. Owing to their crucial role in signal transduction, these proteins are potential drug targets. At present more than 50% of drugs available on the market act through GPCRs. The three-dimensional structures of GPCRs are largely unsolved, except for that of one GPCR (bovine rhodopsin). In contrast, the amino acid sequences of more than 1000 GPCR-related proteins are known (5). Currently known GPCRs include the rhodopsin-like family, the secretin-like receptor family, the metamorphic glutamate receptor-like family, the fungal pheromones mating factor families and cAMP-type receptors (3). The rhodopsin-like family of GPCRs is made up of 15 major subfamilies and more than 60 types of receptor. These receptors bind to diverse ligands and evoke different effector systems. Owing to the enormous amount of data on, and the paramount importance of, GPCRs, an automated computational method for their classification is of great practical use.

In the past, a number of strategies have been used to search for novel GPCRs in protein sequence data. These strategies have involved similarity searches using primary database search tools (e.g. BLAST, FASTA) and such database searches coupled with searches of pattern databases (PRINTS) (6). However, these methods fail when query proteins lack significant sequence similarity to the database sequences. In order to overcome these limitations, a support vector machine (SVM)-based method was developed by Karchin *et al.* (5). This method is better at classifying the subfamilies of GPCRs than simple BLAST or hidden markov model (HMM)-based methods. Another method for classification of eukaryotic

*To whom correspondence should be addressed. Tel: +91 172 2690557, 2690225; Fax: +91 172 2690632, 2690585; Email: raghava@imtech.res.in

GPCRs has been designed based on their membrane spanning topology (7).

This paper describes a support vector machine-based method, GPCRpred, developed for annotating GPCRs on the basis of dipeptide composition. The method uses a three-step approach for annotating GPCRs: (i) it predicts whether the query sequence belongs to the GPCR superfamily or not; (ii) it predicts the class or family of GPCR; and (iii) it predicts the GPCR subfamily if it belongs to Class A of GPCRs. The performance of our method and existing methods was also evaluated on an independent/blind dataset created in this study. Based on the above approach, an online web tool has been developed, which is available at http://www.imtech.res.in/raghava/gpcrpred/.

METHOD

In this study, we have adopted a three-step strategy for recognizing GPCRs from protein sequences and further classifying GPCRs to subfamily level, as shown in Figure 1. The method was trained using fixed-length vectors obtained on the basis of the dipeptide composition of proteins. The accuracy of each step was evaluated using cross-validation. The source of data and strategy used in the development of each module of this method are briefly discussed below.

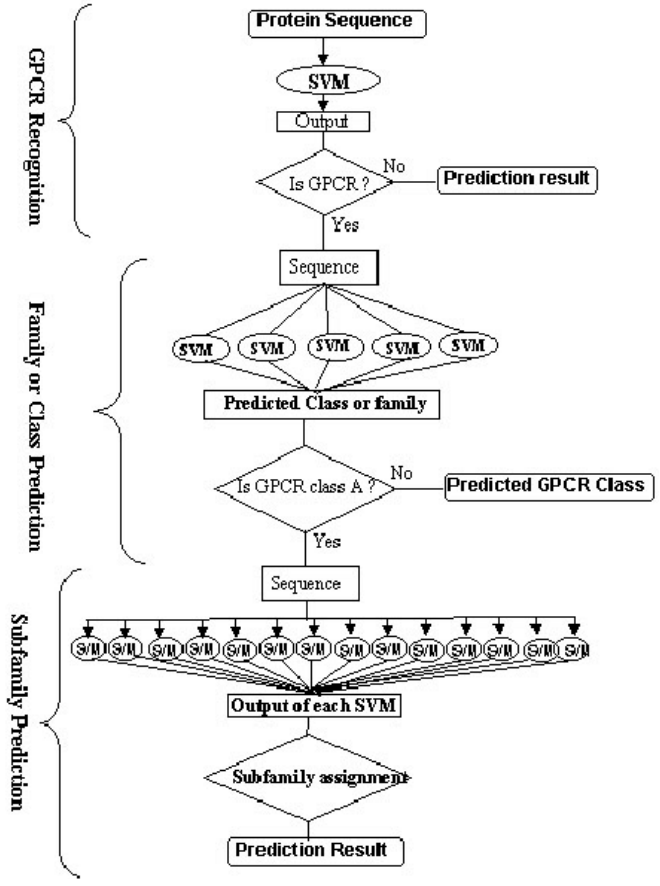


Figure 1. Diagrammatic view of the three-step strategy used to predict the subfamilies of GPCRs.

Recognition of a GPCR

Initially, we developed an SVM module for identifying GPCRs from protein sequence data uncovered by various genome-sequencing projects. The dataset, obtained from http://www.soe.ucsc.edu/research/compbio/gpcr/ (5), consisted of 778 GPCRs belonging to the five major classes of GPCR. This dataset was originally derived from GPCRDB (2). The same dataset was used by Karchin *et al*. for devising a method for the prediction of GPCR subfamilies (5). The dataset was extended by adding 99 decoy negative examples and 2425 additional negative examples obtained from SCOP version 1.37 PDB90 domain data (5,8). The performance of the module was evaluated using a 5-fold cross-validation test. The SVM was trained with a fixed-dimensions vector (400) obtained on the basis of the dipeptide composition of protein sequences.

Recognition of GPCR class

GPCRs can be divided into five major classes: Class A (receptors related to rhodopsin and andrenergic receptors), Class B [receptors related to calcitonin and parathyroid hormone (PTH) receptors], Class C (receptors related to metabotropic receptors), Class D (receptors related to pheromone receptors) and Class E (receptors related to cAMP receptors) (5). The dataset for these five classes was obtained from the work of Karchin *et al*., (5). The dataset consisted of 692 sequences from Class A, 56 sequences from Class B, 16 sequences from Class C, 11 sequences from Class D and 3 sequences from Class E. Classification of GPCRs into one of the five classes is a multi-class classification problem. For five-class classification, five SVMs were constructed, each specific to one class. The *i*-th SVM was trained with all the samples of the *i*-th class with positive labels and samples from the remainder of the classes with negative labels. For example, the SVM for Class A was trained with all Class A sequences with positive labels and the sequences from the other classes (B–E) with negative labels. An unknown GPCR will be classified into the class that corresponds to the SVM with the highest output. The SVM was constructed using the dipeptide composition of proteins. The performance of each SVM was evaluated using 2-fold cross-validation. This 2-fold cross-validation was used because of the lesser number of sequences for Classes D and E.

Non-redundant Dataset

The use of a non-redundant dataset for testing and training is common practice in the development of computational methods to avoid overtraining. Therefore, we generated a non-redundant dataset from the original dataset using PROSET software (9). In the non-redundant dataset, no two sequences had >90% identity. Two SVM modules were constructed at this level, one using the original dataset and the other using the non-redundant dataset.

GPCR subfamily recognition

To predict the subfamily of receptors belonging to the Class A or rhodopsin-like family of GPCRs, we constructed an SVM-based module. Class A was considered because it covers more than 80% of sequences in the GPCRDB database. The Class A family of GPCRs consists of 15 major subfamilies, such as amine, peptide and rhodopsin. The data for all these

subfamilies were extracted from the work of Karchin *et al*. (5). The number of sequences in the different subfamilies of ClassA is provided in Table_sup 3 of the Supplementary Material. The classification of an unknown GPCR into a particular subfamily is a multiclass problem. We constructed K binary SVM classifiers for K subfamily classifications. The i-th SVM was trained using all sequences of the i-th subfamily with positive labels and the sequences from the other subfamilies with negative labels. SVMs trained in this way are referred to as 'one-versus-rest' SVMs (10). The performance of all one-versus-rest SVMs was evaluated using 2-fold cross-validation. An unknown GPCR will be classified into the subfamily that corresponds to the one-versus-rest SVM with the highest output.

Support vector machine

The SVM was implemented using the freely downloadable software package SVM_light written by Joachims (11). The software enables the user to define a number of parameters as well as to select from a choice of inbuilt kernel functions, including a radial basis function (RBF) and a polynomial kernel (of given degree). All the kernel parameters were kept constant except for regulatory parameter C. The experimentation was conducted using various types of kernel such as polynomial and RBF. The SVM was provided with fixed-length vector input. The fixed-length feature vector was obtained from proteins of variable length using dipeptide composition.

Dipeptide composition

The dipeptide composition used as input provides global information on protein features in the form of a fixed-length vector. Dipeptide composition encapsulates information about the fraction of amino acids as well as their local order. The dipeptide composition of each protein was calculated using Equation 1.

$$\text{Fraction of dep}(i) = \frac{\text{Total number of dep}(i)}{\text{Total number all possible dipeptides}}, \quad \mathbf{1}$$

where dep(i) is a dipeptide i out of 400 dipeptides.

In this study, 20 SVMs were constructed in total: one for discriminating GPCR proteins from other proteins such as globular proteins, five for predicting the class or family of GPCRs and the remainder for recognizing the subfamily of a GPCR that belongs to Class A.

Independent or blind dataset

In this study, an independent/blind dataset was created for unbiased evaluation of this as well as existing methods (5). GPCR sequences were derived from release 8.0 of GPCRDB (2). All sequences previously used by our methods and existing methods for training or testing were removed from the dataset. All proteins denoted as 'fragment' or whose annotation was listed as 'hypothetical', 'similar' or 'putative' were also removed from the dataset. The final dataset comprises 650 proteins belonging to five major classes of GPCRs, as shown in Table 4. The dataset had 431 sequences of Class A belonging to 10 major subfamilies (Table 5).

Performance evaluation

The performance of SVM in distinguishing GPCRs from non-GPCRs was evaluated using 5-fold cross-validation. In the 5-fold cross-validation, the dataset was partitioned randomly into five equal-sized sets. The training and testing of each classifier was carried out five times using one distinct set for testing and other four sets for training. Four threshold-dependent parameters—sensitivity, specificity, accuracy and Matthews's correlation coefficient (MCC)—were used to measure the performance of this module. The performance of SVM modules constructed for recognizing GPCR class and subfamily was evaluated using 2-fold cross-validation because of the lower number of sequences. In the 2-fold cross-validation, the dataset was randomly partitioned into two equal sets. The training and testing of each SVM was carried out twice using one set for training and the other set for testing. The performance of these modules was measured using two parameters, accuracy and MCC, as described by Hau and Sun (2001) for prediction of a protein's subcellular localization (10).

RESULT AND DISCUSSION

The three-step strategy used for the development of our method is shown in Figure 1. The input for all SVMs is a fixed-length vector obtained using dipeptide composition from the primary amino acid sequence. The performance of each SVM module was validated using a cross-validation test.

The performance of the module developed for discriminating between GPCRs and other protein sequences is summarized in Table 1. The results depict that the method can differentiate GPCRs from other proteins with an accuracy of 99.5% and an MCC of 0.99 at a default cutoff score of 0, when evaluated through 5-fold cross-validation. The best results were obtained using the RBF kernel with $\gamma = 200$. The value of regulatory parameter C was optimized to 100. Another SVM-based module was developed using 2872 negative protein sequences with 778 GPCR sequences. These 2872 protein sequences were retrieved from the EVA server (http://cubic.boic.coulmbia.edu/eva/res/week.html#unique) on November 25, 2002. The performance of the module developed using the EVA dataset is shown in Table_sup1 (see

Table 1. The performance of our method in differentiating GPCRs from non-GPCRs

Threshold	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
−1.0	99.9	87.8	90.6	0.74
−0.8	99.9	94.0	95.4	0.89
−0.6	99.8	97.1	97.8	0.94
−0.4	99.3	98.8	98.9	0.97
−0.2	99.2	99.5	99.5	0.99
0.0	98.6	99.8	99.5	0.99
0.2	96.5	99.9	99.0	0.98
0.4	93.3	100.0	98.4	0.95
0.6	86.4	100.0	96.8	0.91
0.8	74.8	100.0	94.0	0.83
1.0	53.3	100.0	88.9	0.67

The highlighted region specifies the performance of the method at the default threshold.

Supplementary Material). The performance of this module is very close to the performance of the SVM module developed using the dataset suggested by Karchin *et al*. (5). The results demonstrated that GPCRs and non-GPCRs can be classified with high accuracy using dipeptide composition as the sequence feature.

The performance of the dipeptide composition-based module was compared with that of an amino acid composition-based module developed on the same dataset. The performance of the amino acid composition-based module is shown in Table_sup2 (http://www.imtech.res.in/raghava/info.html). The module was able to discriminate between GPCRs and non-GPCRs with an accuracy of 96.5% at the default cutoff score, which was lower than that of the dipeptide composition-based module. This proves that dipeptide composition is a better feature for recognizing GPCRs from protein sequence data. These results support our previous observation that dipeptide composition is a better feature for predicting the subcellular localization of proteins (12). The performance of the dipeptide composition-based module was also compared with a similarity search-based tool (BLAST). In the case of BLAST (at an *E*-value cutoff of 10^{-10}) during 5-fold cross-validation no significant hit was obtained for 102 GPCR proteins out of 778 proteins. The accuracy and MCC of the BLAST-based module were 86.5% and 0.91, which were significantly lower than the figures for the dipeptide composition SVM-based module. These observations suggest that it is worth using the computationally expensive SVM-based tools to recognize GPCRs from the genomic data produced through ongoing sequencing projects such as Human Genome.

To predict the class or family of GPCRs, a series of binary SVMs were constructed. The SVMs were trained and tested using dipeptide compositions through 2-fold cross-validation. The performance of the SVMs in recognizing the classes or families of GPCRs is summarized in Table 2. The results show that the SVM-based method is able to discriminate between all the classes with more than 80% accuracy, except Class D. Poor results were obtained for Class D owing to the lower number of sequences for training. It is a well-established fact that learning techniques require a large number of examples for reliable prediction. The overall accuracy and MCC of this module for predicting the five GPCR classes were 97.3% and 0.81 respectively. In order to examine the effect of similar sequences in the dataset on the performance of the method, the performance of the method on the non-redundant dataset was also evaluated. The performance of the method developed using the non-redundant dataset is shown in Table 2. The overall performance of this method was very close to the performance of the method developed using the original dataset. This observation proves that our method can perform well on both similar and diverse sequences.

The comparison of a newly developed method with existing methods is necessary to instil confidence in users. Towards this end, we have compared the performance of our methods with the existing GPCR subfamilies classification method developed by Karchin *et al*. (5). The performance was compared by computing the GPCRs correctly predicted before the first false positive error. Our method correctly predicted 98% of GPCRs before the first error was observed, which is similar to the performance of the HMM-based method developed by Karchin *et al*. (5). The performance of the dipeptide composition-based method was nearly 13% greater than the existing Fisher score vector (FSV)-based method for recognizing the five classes of GPCR (5). This proves that dipeptide composition is a better feature in terms of encapsulating global information about proteins. Dipeptide composition provides information about the fraction of amino acids as well as the local order of the amino acids. In the past, it has been observed that dipeptide composition can classify proteins with superior accuracy than amino acid composition or pseudo-amino acid composition (12–15).

The identification of GPCR subfamilies is of major interest to pharmaceutical companies and experimental biologists. The prediction of subfamilies is crucial in assigning a function to GPCRs. Therefore, we have developed modules for classifying the subfamilies of the rhodopsin-like family. The performance of this module was evaluated using 2-fold cross-validation. The performance of the module in predicting the subfamilies in terms of accuracy and MCC is shown in Table 3. The prediction accuracy for most of the subfamilies was >85%. The overall accuracy and MCC for 14 subfamilies of the rhodopsin-like family of GPCRs were 97.3% and 0.97 respectively. Poorer results were achieved for a few subfamilies owing to the lower number of sequences for training. The results suggest that the different subfamilies of the rhodopsin-like family are quite closely correlated with dipeptide composition, implying that GPCR subfamilies are predictable with superior accuracy if a good training dataset can be established for this purpose.

Performance on the independent dataset

It is ideal to evaluate methods on an independent/blind dataset to demonstrate their true or unbiased performance. The sequences of the independent dataset were used neither for

Table 2. The performance of our method in identifying the five major classes of GPCR

GPCR families (Class)	Original			Non-redundant		
	Seq	ACC (%)	MCC	Seq	ACC (%)	MCC
Rhodopsin and andrenergic-like receptors (Class A)	692	98.1	0.80	496	98.9	0.86
Calcitonin and PTH-like receptors (Class B)	56	85.7	0.84	40	95.0	0.96
Metabotropic-like receptors (Class C)	16	81.3	0.81	12	100.0	0.86
Pheromone-like receptors (Class D)	12	36.4	0.49	11	36.6	0.60
cAMP-like receptors (Class E)	3	100.0	1.00	3	33.3	0.58

Seq, ACC and MCC: number of sequences, accuracy and Matthew's correlation coefficient, respectively. The non-redundant dataset was created from the original dataset after removing sequences having >90% identity with other sequences in the dataset.

training nor for testing during the development of the method. Therefore, in this study, the performance of our method and an existing method (5) was evaluated at class level as well as at subfamily level using the independent dataset (Table 4). The method developed in this study and Karchin *et al*.'s method were able to correctly predict 593 (91.2%) and 542 (83.3%) GPCRs, respectively (Table 4). This proves that the performance (91.2%) of our method is superior to that of the existing method. These results also demonstrate that the performance of both methods is highly accurate on the independent dataset.

The performance of our method as well as of Karchin *et al*.'s method (5) at subfamily level was also evaluated using this dataset. The performance of our method in predicting the subfamilies of Class A is shown in Table 5. Unfortunately, Karchin *et al*.'s method did not predict any subfamilies due to some technical problem on the server. The overall accuracy of our method for major subfamilies of Class A is ~93%. These results clearly indicate that our method is highly accurate for data not used for training and testing.

We applied our SVM-based method to a set of peptides from GPCRDB that were not classified beyond inclusion in Class A (2). We considered a value of 1.0 for the difference between highest and second highest SVM to assign subfamilies with good reliability. The method is able to assign subfamilies to 42 unclassified protein sequences. The list of protein sequences assigned subfamilies by our method is provided in Table_sup4 of the Supplementary Material. We do not have the resources in our lab to verify our predictions by wet experimentation. Users are encouraged to do so.

This study illustrates a new principle for the classification of GPCRs when information about the features of a protein in the primary sequence is extracted in the form of dipeptides. The dipeptide compositions transduce the protein information to fixed-length vectors, which is a crucial requirement for the development of machine learning techniques-based methods. The most notable feature of our method is its ability to distinguish all the families as well as the subfamilies of GPCRs with extremely high accuracy. Thus, this method can be used to recognize novel GPCRs as well as their functional classification. The method can assist in automated functional annotation of genomic data and can help in reducing the gap between the amount of genomic sequence data produced and the annotation rate.

GPCRpred SERVER

Based on our study, we have developed a web server that allows the user to recognize and classify GPCRs from primary amino acid sequences. GPCRpred is freely available at http://www.imtech.res.in/raghava/gpcrpred/. The common gateway interface (CGI) script for GPCRpred is written using PERL version 5.03. This server is installed on a Sun Server (420E) under a UNIX (Solaris 7) environment. Users can enter the primary amino sequence in any standard format (EMBL/FASTA/GCG) or plain text format. The server uses the

Table 3. The performance of our method in classifying the major subfamilies of the rhodopsin and andrenergic family of GPCRs

GPCR Class A subfamilies	ACC (%)	MCC
Amine	99.1	0.99
Peptide	99.7	0.95
Hormone proteins	100.0	1.00
Rhodopsin	98.9	0.99
Olfactory	100.0	0.99
Prostanoid	100.0	0.99
Nucleotide-like	85.4	0.92
Cannabis	100.0	1.00
Platelet activating factor	100.0	1.00
Gonadotrophin releasing hormone	100.0	1.00
Thyrotropin releasing hormone	85.7	0.93
Melatonin	100.0	1.00
Viral	33.3	0.58
Lysospingolipids	58.8	0.76
Overall	97.3	0.97

ACC and MCC: accuracy and Matthew's correlation coefficient, respectively.

Table 5. The performance of our method on a independent dataset in recognizing 10 major subfamilies of the rhodopsin-like family

GPCR Class A subfamilies	Total	Correctly predicted
Amine	25	24
Peptide	69	68
Hormone proteins	2	2
Rhodopsin	7	5
Olfactory	296	270
Prostanoid	3	3
Nucleotide-like	7	6
Cannabis	1	1
Gonadotrophin releasing hormone	10	9
Lysospingolipids	11	11
Overall	431	399 (92.6%)

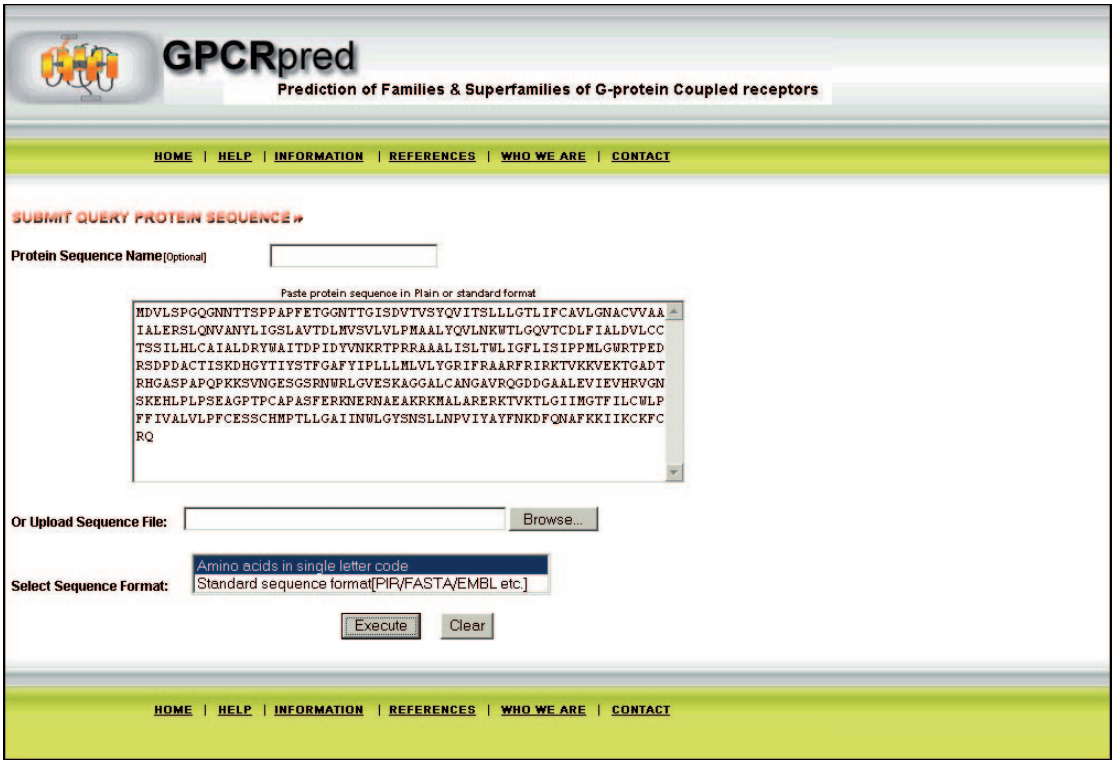
The numbers in brackets specify the overall accuracy of method.

Table 4. The performance of our method and Karchin *et al*.'s method (5) on an independent dataset at class level

GPCR families (Class)	Total sequences	Correctly predicted sequences	
		Our method	Karchin *et al*.'s method
Rhodopsin and andrenergic-like receptors (Class A)	431	431	427
Calcitonin and PTH-like receptors (Class B)	129	111	20
Metabotropic-like receptors (Class C)	76	43	76
Pheromone-like receptors (Class D)	12	7	12
cAMP-like receptors (Class E)	2	1	2
Overall	650	593 (91.2%)	542 (83.3%)

The numbers in brackets specify the overall accuracy of methods.

(a)



(b)

Figure 2. The GPCRpred home page and result page. (**A**) The GPCRpred homepage showing the principle features of the interface. (**B**) A GPCRpred results page showing a summary of the submitted sequence and final prediction results.

ReadSeq subroutine developed by Don Gilbert to parse the input sequence. Users can submit sequences for prediction using file uploading or cut-and-paste options, as illustrated in Figure 2A. After analysis, the results are shown in a user-friendly format. The results provide summarized information about the query sequence and prediction. An example of prediction output is shown in Figure 2B.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to the Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India for financial assistance. The manuscript has IMTECH communication no. 015/2004.

REFERENCES

1. 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.
2. Horn,F., Vriend,G. and Cohen,F.E. (2001) Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. *Nucleic Acids Res.*, **29**, 346–349.
3. Attwood,T.K., Croning,M.D. and Gaulton,A. (2002) Deriving structural and functional insights from a ligand-based hierarchical classification of G protein-coupled receptors. *Protein Eng.*, **15**, 7–12.
4. Sadowski,M.I. and Parish,J.H. (2003) Automated generation and refinement of protein signatures: case study with G-protein coupled receptors. *Bioinformatics*. **19**, 727–734.
5. Karchin,R., Karplus,K. and Haussler,D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, **18**, 147–159.
6. Lapinsh,M., Gutcaits,A., Prusis,P., Post,C., Lundstedt,T. and Wikberg,J.E. (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.
7. Inoue,Y., Ikeda,M. and Shimizu,T. (2004) Proteome-wide functional classification and identification of mammalian-type GPCRs by binary topology pattern. *Comp. Biol. Chem.*, **28**, 39–49.
8. Jaakkota,T., Diekhans,M. and Haussler,D. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.*, **7**, 95–114.
9. Brendel,V. (1992) PROSET—a fast procedure to create non-redundant sets of protein sequences. *Mathl. Comput. Modelling*, **16**, 37–43.
10. Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.
11. Joachims,T. (1999) Making large-scale SVM learning practical. In Scholkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods Support Vector Learning*. MIT Press, Cambridge, MA and London, pp. 42–56.
12. Bhasin,M. and Raghava,G.P.S. (2004) ESLPred: SVM based prediction of subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.*, **32**, W414–W419.
13. Bhasin,M. and Raghava,G.P.S. (2004) Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.* (in press).
14. Grassmann,J., Reczko,M., Suhai,S. and Edler,L. (1999) Protein fold class prediction: new methods of statistical classification. In Lengauer,T., Schneider,R., Bork,P., Brutlag,D.L., Glasgow,J.I., Mewes,H.-W. and Zimmer, R. (eds), *Proceeding of Seventh International Conference on Intelligent System for Molecular Biology (ISMB'99)*, AAAI, Heidelberg, Germany, pp. 106–112.
15. Reczko,M. and Bohr,H. (1995) The DEF database of sequence based protein fold class prediction. *Nucleic Acids Res.*, **22**, 3616–3619.