Bioinformation
www.bioinformation.net
open access
Hypothesis

A model for the evaluation of domain based classification of GPCR

Tannu Kumari*, Bhaskar Pant, Kamalraj Raj Pardasani

Department of Mathematics, MANIT, Bhopal - 462051, India; Tannu Kumari – Email: ttannu@gmail.com; *Corresponding author

Received July 08, 2009; Revised July 30, 2009; Accepted September 11, 2009; Published October 11, 2009

Abstract:
G-Protein Coupled Receptors (GPCR) are the largest family of membrane bound receptor and plays a vital role in various biological processes with their amenability to drug intervention. They are the spotlight for the pharmaceutical industry. Experimental methods are both time consuming and expensive so there is need to develop a computational approach for classification to expedite the drug discovery process. In the present study domain based classification model has been developed by employing and evaluating various machine learning approaches like Bagging, J48, Bayes net, and Naive Bayes. Various softwares are available for predicting domains. The result and accuracy of output for the same input varies for these software's. Thus, there is dilemma in choosing any one of it. To address this problem, a simulation model has been developed using well known five softwares for domain prediction to explore the best predicted result with maximum accuracy. The classifier is developed for classification up to 3 levels for class A. An accuracy of 98.59% by Naïve Bayes for level I, 92.07% by J48 for level II and 82.14% by Bagging for level III has been achieved.

Keywords: GPCR; model; membrane proteins; Bayes model

Background:
G-Protein-Coupled-receptors (GPCR) are the largest family of membrane bound receptor and they play a significant role in mediating various biological processes. They have seven hydrophobic regions that cross the membrane, an amino terminal region outside the cell, 3 intracellular loops, 3 extracellular loops followed by C terminal region in intracellular region (Figure 1a). A diverse array of chemical substances act as ligand, including amino acid, ions, lipids, peptide hormones, chemokines, odorants, hormones, pheromones, odorants, purines, neuropeptides, tastants **[1].** GPCR are considered as an excellent potential therapeutics target class for drug design and the focus of current pharmaceutical research and therapeutic intervention. Traditional experimental method are very expensive and time consuming so there is need to develop computational models to expedite the drug discovery process.

Domains are considered to be the molecular signatures. They are the building blocks of proteins. A protein domain is a structurally compact, independently folding unit that forms a stable three-dimensional structure and shows a certain level of evolutionary conservation. Typically, a conserved domain contains one or more motifs. During evolution, they have been duplicated, fused and recombined, to produce proteins with novel structures and functions. Domain varies in length between 25 amino acids up to 500 amino acids. One domain may appear in a variety of evolutionary related proteins. Each domain forms a compact three dimensional structure and often can be independently stable and folded. Some protein domains are "promiscuous" and can be found in association with a variety of other domains. Therefore, during protein sequence analysis, it is often advantageous to deal with one domain at a time. The shortest domains such as zinc fingers are stabilized by metal ions or disulphide bridges. Domains often form functional units, such as calcium-binding, EF domain etc.

Attempts have been made by various research groups to develop classifiers. The first classification attempt was made by Attwood and Findlay, when they developed sequence based finger prints of the seven characteristics GPCR hydrophobic domains. Kolakowski gave the important, well known A-F classification system **[2].** Bockaert & Pin represented classification system on the basis of structural and ligand binding criteria classified GPCR in five classes **[3].** After the availability of human genome in 2001, Fredriksson and colleagues, **[4]** classified GPCR in five major classes commonly known as "GRAFS" (glutamate, rhodopsin, adhesion, frizzled and secretin) based on phylogenetic criteria. Elrod and Chou, **[5]** suggested a covariant discriminant algorithm to predict GPCR's type from amino acid composition. Karchin *et al,* **[6]** developed a system based on support vector machines built on profile HMMs. Inoue *et al.,* **[7]** gave the classification by binary topology pattern. Qian *et al*., **[8]** suggested a phylogenetic tree based profile Hidden Markov model (T-HMM) for GPCR classification. Papasaikas *et al.,* **[9]** developed classification using sequence alone using signatures derived from profile hidden markov models. Gaulton and Attwood **[10]** used bioinformatics approaches for GPCR classification. Kuo-Chen Chou, **[11]** generated model based on primary sequence by using covariant discriminant predictor. Yang and Deogan, **[12]** used probabilistic suffix tree prediction model for each of the subfamilies. Erguner *et al.* **[13]** developed a classification model based on ligand specific features. Mathew N. Davies *el al*., **[14]** performed classification based on sequence and motif. Gangal and Kumar, **[15]** made classification based on reduced alphabet motif methodology. Mathew *et al.* **[14]** performed classification based on simple representation of a protein's physical properties. Gupta *et al.* **[16]** performed classification on dipeptide based SVM approach. But from the literature survey it appears that no attempt has been made to develop computational approaches for the classification of GPCR using domains. Thus, an attempt has been made to develop a model for domain based classification of GPCRs.

Methodology:
GPCR superfamily can be organized into a hierarchy of classes, class sub-families and class sub-sub-families according to GPCR database (GPCRDB) **[17]**. Here the GPCR family tree has been shown, of which the present work focuses on the further classification of Amine subfamily up to sub-sub-family level. Various softwares are available for prediction of domains which have been developed using different approaches such as SVM, HMM, Neural Network etc. Thus, for the same input they give different result and also differ in accuracy. This variation in result and accuracy leads to dilemma of choosing software for prediction of domains. This information of domains is required in the proposed classifier. Classification using merely the predicted domain from the input sequence. Here five well known softwares, namely SBASE (SB), SMART (SM), NCBI CONSERVED DOMIAN (NC), SCAN PROSITE (SC), and PHYLODOME (PHY) have been used. Sbase, a support vector machine based tool, is a collection of protein domain sequences collected from the literature, protein sequence databases and genomic databases. The protein domains are defined by their sequence boundaries given in one of the primary sequence databases (Swiss-Prot, PIR, TREMBL etc.) **[18]**. Smart, simple modular architecture research tool is a web based tool that allows domain identification and annotation. The tool compares every sequence with its databases of domain sequences and multiple alignments as well as identifies compositionally biased regions such as signal peptide, transmembrane and coiled coil segments **[19]**.

NCBI conserved domain database (CDD) is a collection of multiple sequence alignments and derived database search models, which

represent protein domains conserved in molecular evolution. CDD provides annotation of domain footprints and conserved functional sites on protein sequences **[20]**. Scan PROSITE is a new and improved version of the web-based tool for detecting PROSITE signature matches in protein sequences. For a number of PROSITE profiles, the tool uses ProRules to detect functional and structural intra-domain residues **[21]**. Phylodome performs the analysis of taxonomic distribution and lineage-specific variation of domains and domain combinations. It provides a fast overview on the taxonomic spreading and potential interrelation of domains. PhyloDome is a tool which can visualize and analyze the phylogenetic distribution of one or more eukaryotic domains **[22]**. Different Machine learning approaches such as Naïve Bayes, Bayes Net, J48 and Bagging of WEKA has been employed **[23]**.

Naive bayes classifier: It is a simple probabilistic classifier based on applying Bayes theorem with strong independent assumptions. In other words, Naïve Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variable. This assumption is called class conditional independence. It is made to simplify the computation and in this sense considered to be "Naïve". The conditional independence assumption can be formally stated as in equation 1 (see supplementary material):

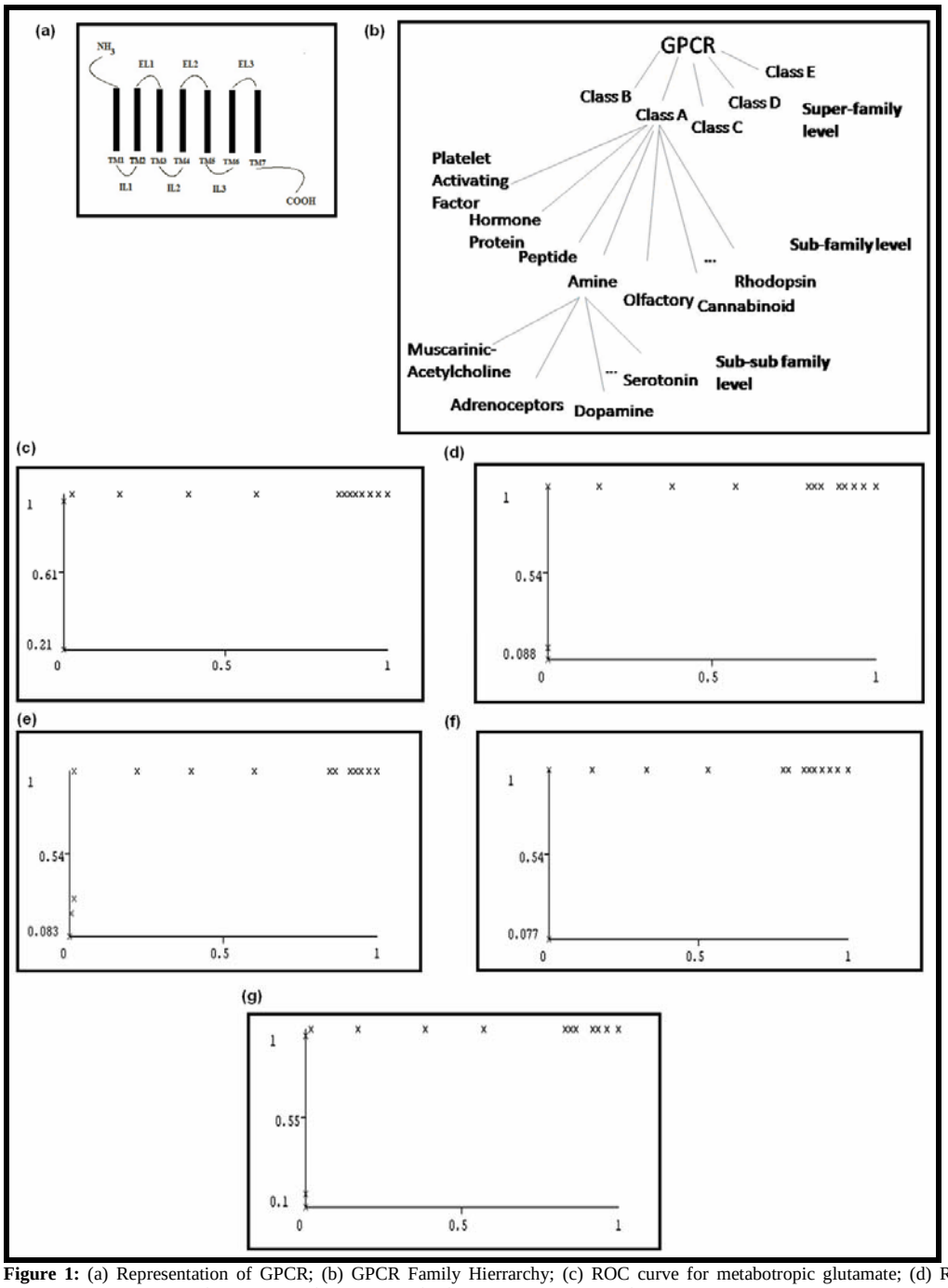

Figure 1: (a) Representation of GPCR; (b) GPCR Family Hierrarchy; (c) ROC curve for metabotropic glutamate; (d) ROC curve for rhodopsin; (e) ROC curve for adhesion; (f) ROC curve for frizzled; (g) ROC curve for secretin

Bayes net: It represents a more flexible approach for modelling the class conditional probabilities P(X/Y). This approach instead of requiring all the attributes to be conditionally independent specifies the exact pair of attributes that are conditionally independent **[23]**.

J48: A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals. The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery **[23]**.

Bagging: Bagging also called as bootstrap aggregating, is a technique that repeatedly samples from a data set according to a uniform probability distribution. Each bootstrap sample has the same size as the original data **[23]**.

The proteins used for this study were collected from GPCRDB (G-Protein Coupled Receptor Database) **[17]**. The sequences of Protein in GPCRDB are derived from the SWISS-PROT and TREMBL Data banks. The incomplete sequences containing fragments were removed. NRDB program was used to verify that none of the sequences were identical to each other in the data set.

Discussion:

Receiver Operating Curve (ROC) is a graphical technique for evaluating data mining schemes. ROC curves depicts the performance of a classifier without regard to class distribution or error costs .They plot the number of positives included in the samples on the vertical axis, expressed as a percentage of the total number of positives, against the total number of negatives on the horizontal axis. For each fold of a 10 fold cross validation ,weight the instances for a selection of different cost ratios train the scheme on each weighted set, count the true positives and false positives in the test set, and plot the resulting point on the ROC axes. The ROC curves for different classes have been plotted as shown in **Figures (1b-g)**. As ROC depicts the performance, we can refer from the confusion matrix that in case of Metabotropic Glutamate class, the false positive ratio is 0, which clearly indicates that the true positive ratio is 100% i.e. 1. Similarly with Rhodopsin, frizzled and Secretin the ratio is 1 .In case of Adhesion, the false positive value is 0.017, which shows that the ratio is below 1 and it is 0.986. The accuracy of results for the three levels obtained from all the four classifiers with input as domains predicted from five different softwares and their combinations is presented in (**Tables 1-8 in supplementary material**) given in appendix.

In the FAMILY LEVEL (**see Table 1 in supplementary material**), when predicted domains from Sbase are taken, the accuracy of is 98.59% is achieved, which is consistent with all classifiers used and is highest as compared to results obtained with input predicted from remaining four softwares individually. Further with two, three and four combinations of softwares, involving Sbase gives same accuracy of 98.59%. This depicts that all the domains that are predicted by the other softwares are also predicted by Sbase only, as Sbase predicts more number of domains than the remaining four softwares used. In the SUB FAMILY LEVEL, when input is taken as domains predicted by Sbase, it gives accuracy of 90.85% which is highest as compared to the results obtained by classifiers with input predicted from remaining four softwares (**see Table 2 in supplementary material**) individually. But when Sbase is used in combination with Smart to predict domains as input to J48 classifier, it gives 92.07 % accuracy (**see Table 3 in supplementary material**). This implies that adding input from Smart improves accuracy marginally from 90.85- 92.07%. Further with three and four combinations, same accuracy is achieved, indicating no improvement in accuracy due to redundancy in prediction of domains. In the SUB-SUB FAMILY LEVEL, domains predicted by Sbase as input to all classifier gives an accuracy of 80.35 % consistently (**see Table 4 in supplementary material**). For combinations of two softwares for predicting domains as input to classifiers gives an accuracy of 82.14% with bagging (**see Table 5 in supplementary material**). Similarly in this level too, with three and four combinations the highest accuracy obtained is 82.14%. The accuracy of results obtained is comparable with those obtained by earlier research workers **[18]** and shown below:

Analysis:

Among all the five softwares, the domains predicted by Sbase when used as input to all the classifiers trained, consistently give result s with best accuracy. Thus we conclude that Sbase predicts domains with better accuracy compared to remaining four software used. Further it is concluded that no single classifier works best for all the three levels. Hence classifier Bagging for level III, J48 for level II and all the four classifiers (J48, Bagging, Naïve Bayes and Bayes Net) for level I are recommended to achieve better accuracy.

Acknowledgement:

The authors are highly thankful to Department of biotechnology, New Delhi for providing Bioinformatics Infra Structures Facility at MANIT, Bhopal for carrying out this work.

References:

[1] TK Attwood, JB Findlay, *Protein Eng.* **6**: 167 (1993) [PMID: 8386361]
[2] Jr. LF Kolakowski *Receptors channels* **2**: 1 (1994) [PMID: 8081729]
[3] J Bockaert, JP Pin, *EMBO J.* **18**: 1723 (1999) [PMID: 10202136]
[4] R Fredriksson *et al., Mol Pharmacol,* **63:** 1256 (2003) [PMID: 12761335]
[5] D W Elrod, KC Chou, *J Proteome Res.* **1**: 429 (2002) [PMID: 12645914]
[6] R Karchin *et al., Bioinformatics* **18**: 147 (2002) [PMID: 11836223]
[7] Y Inoue *et al., Computational biology and Chemistry* 28: 39 (2004) [PMID: 15022640]
[8] B Qian *et al., FEBS Lett.* **554:** 95 (2003) [PMID: 14596921]
[9] PK Papasaikas *et al., Nucleic Acids Res.* **32**: W380 (2004) [PMID: 15215415]
[10] A Gaulton, TK Attwood, *Nucleic Acids Res.* **31**: 3333 (2003) [PMID: 12824320]
[11] KC Chou *Journal of Proteome Research* **4**: 1413 (2005) [PMID: 16083294]
[12] L Goodstadt, CP Ponting, *Plos Comp Biol* 2:e1333 (2004) [PMID: 17009864]
[13] Y Okuno *et al., Nucleic Acid Res.* **34:**D673 [PMID: 16381956]
[14] MN Davies *et al., Bioinformatics* **23**: 3113 (2007) [PMID: 17956878]
[15] R Gangal, KK Kumar, *J Biomolecular Structure Dynamics* **25**: 299 (2007) [PMID: 17937491]
[16] R Gupta *et al., IEEE Transactions on information Technology in Biomedicine* **12**: 541 (2008) [PMID: 18632334]
[17] http://www.gpcr.org/7tm/
[18] http://hydra.icgeb.trieste.it/sbase/
[19] http://smart.embl-heidelberg.de/
[20] http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
[21] http://www.expasy.ch/tools/scanprosite/
[22] http://mendel.imp.ac.at/PhyloDome/
[23] http://www.cs.waikato.ac.nz/~ml/weka/
[24] Y Huang *et al., Computational biology and chemistry,* **28**: 275 (2005) [PMID: 15548454]

Edited by P. Kangueane
Citation: Kumari *et al.,* Bioinformation 4(4): 138-142 (2009)

Supplementary material:

Equation 1

$$P(X/Y=y)=\prod_{i=1}^{d} P(X_i/Y=y)$$

Where each attribute set $X=\{X_1, X_{2...} X_d\}$ consists of *d* attributes **[23]**.

Table 1: Comparison of Results

	Family level	Sub-family level	Sub-Sub Family level
Present work	98.59 %	92.07%	82.14%
Y. Huang *et. al* **[24]**	Not available	91.1%	82.4%
Mathew *et. al.* **[14]**	97%	84%	75%

Table 2: Accuracy of classification of FAMILY (level 1)

SOFTWARE →	SBASE	SMART	NCBI CONSERVED	SCAN PROSITE	PHYLODOME
CLASSIFIER ↓					
J48	98.59	71.83	95.07	96.47	97.18
BAYES NET	98.59	71.83	95.07	96.47	97.18
NAÏVE BAYES	98.59	71.83	95.07	96.47	97.18
BAGGING	98.59	71.83	95.07	96.47	97.18

Table 3: Accuracy of classification of SUB-FAMILY (level 2)

SOFTWARE →	SBASE	SMART	NCBI CONSERVED	SCAN PROSITE	PHYLODOME
CLASSIFIER ↓					
J48	90.85	28.65	31.09	29.87	26.21
BAYES NET	90.85	28.65	31.09	29.87	26.21
NAÏVE BAYES	90.85	28.65	31.09	29.87	26.21
BAGGING	90.85	28.04	31.09	29.87	26.21

Table 4: Accuracy of classification of SUB-FAMILY (level 2) with 2 combinations

SOFTWARE →	SB+PHY	SB+SC	SB+SM	SB+NC	SB+SC	PHY+SM	PHY+NC	SC+SM	SC+NC	SM+NC
CLASSIFIER ↓										
J48	90.85	90.85	92.07	90.85	90.85	31.7	33.53	39.02	38.41	37.19
BAYES NET	90.24	87.80	81.09	90.24	87.80	32.92	33.53	37.8	38.41	37.19
NAÏVE BAYES	80.48	78.04	72.56	79.26	78.04	32.31	33.53	37.8	38.41	37.19
BAGGING	90.85	90.85	90.85	90.85	90.85	30.48	33.53	39.02	38.41	36.58

Table 5: Accuracy of classification of SUB-SUB-FAMILY (level 3)

SOFTWARE →	SBASE	SMART	NCBI CONSERVED	SCAN PROSITE	PHYLODOME
CLASSIFIER ↓					
J48	80.35	33.92	26.78	33.92	26.78
BAYES NET	80.35	33.92	28.57	33.92	30.55
NAÏVE BAYES	80.35	33.92	28.57	33.92	30.55
BAGGING	80.35	33.92	26.78	33.92	28.57

Table 6: Accuracy of classification of SUB-SUB-FAMILY (level 3) for two combinations

SOFTWARE →	SB+PHY	SB+SC	SB+SM	SB+NC	SB+SC	PHY+SM	PHY+NC	SC+SM	SC+NC	SM+NC
CLASSIFIER ↓										
J48	80.35	80.35	80.35	26.78	80.35	37.5	26.78	37.5	33.92	33.92
BAYES NET	80.35	80.35	80.35	78.51	80.35	35.71	30.35	35.71	33.92	33.92
NAÏVE BAYES	78.57	80.35	73.21	78.57	80.35	35.71	30.35	35.71	33.92	33.92
BAGGING	80.35	80.35	82.14	80.35	80.35	35.71	28.57	33.92	33.92	33.92

Table 7: Confusion matrix for different classes by Naive Bayes

a	b	c	d	e	<---classified as
27	0	1	0	0 |	a = mg
0	34	0	0	0 |	b = rhodopsin
0	0	24	0	0 |	c = adhesion
0	0	0	26	0 |	d = frizzled
0	0	1	0	29 |	e = secretin

Table 8: Detailed accuracy by class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC	Area Class
	0.964	0	1	0.964	0.982	1	mg
	1	0	1	1	1	1	rhodopsin
	1	0.017	0.923	1	0.96	0.986	adhesion
	1	0	1	1	1	1	frizzled
	0.967	0	1	0.967	0.983	1	secretin
Weighted Avg.	0.986	0.003	0.987	0.986	0.986	0.998	