Journal of Advanced Bioinformatics Applications and Research ISSN 0976-2604
Vol 2, Issue 2, June-2011, pp 125-129 http://www.bipublication.com

SVM BASED EVOLUTIONARY MODEL FOR CLASSIFICATION OF BACTERIA, FUNGI, ANIMAL AND HIV GPCR'S

Anubha Dubey*, Kumud Pant*, Usha Chouhan**
*Department of Bioinformatics
**Department of Mathematics
MANIT, BHOPAL, 462051
INDIA

ABSTRACT

Classical methods have been recruited to determine the molecular function and the physiological relevance of G-protein-coupled receptors (GPCRs), including ligand binding and signal transduction studies, pharmacological receptor profiling in tissues and the characterization of transgenic mouse models. Evolutionary data from both sequenced genomes and targeted retrieved orthologs are increasingly used as a source of structural information. Recent success in sequencing and functionally expressing GPCRs from fossils opens the possibility of studying signaling pathways even in extinct species. Therefore, mining evolutionary data provides an additional source for understanding the functional relevance of individual GPCRs, for interpreting naturally occurring receptor mutations in patients and for guiding structural modelling and mutagenesis studies of GPCRs. In the present study a machine learning model has been developed to classify GPCR proteins from bacteria, fungi, animal and HIV to see the evolutionary pattern. Hence by evaluating grid search with 5 fold cross validation c and g values of bacteria, fungi, animal and HIV is obtained with accuracies of 99.8626%, 99.0698%, 99.6613%, and 99.60% respectively.

Keywords: Signal pathway, Transgenic, GPCR, Mutagenesis

Background:

The availability of complete genome sequence data for human [1], chimpanzee [2] and many other species has provided the opportunity to compare entire genomes globally. Furthermore, the shared evolutionary ancestry of many proteins has enabled the development of comparative approaches designed to identify conserved sequence motifs that are responsible for certain functions. Therefore, Dobzhansky's famous quote 'nothing in biology makes sense except in the light of evolution' [3] is more relevant than ever. This is also true for research into G-protein coupled receptors (GPCRs). Using genomic data, questions about the evolutionary origin of present-day GPCRs, their signaling pathways, the biological relevance of certain GPCRs over time and the stability of structurally and functionally relevant elements can be addressed with much greater accuracy. Finally, GPCRs constitute the most abundant gene family in most animal genomes. Their large number and highly conserved architecture, combined with a reasonable degree of variability, render GPCRs excellent candidates for studying evolutionary mechanisms at a molecular level.

The conceptual basis of many evolutionary approaches is the fact that the structural diversity of a given receptor protein among different species is the result of a long evolutionary process, which is characterized by a continuous accumulation of mutations. If the receptor is required for the maintenance of vital functions in an organism, sufficient structural conservation is necessary to ensure the functionality of the receptor protein. The occurrence of GPCRs and G-protein signaling dates back 1.2 billion years, which is before plants, fungi and animals emerged from a common ancestor. The phylogenetically oldest GPCRs that have been studied to date include fungal pheromone receptors, cAMP-receptor-like receptors and glutamate-receptor-like receptors. For example, glutamate-receptor-like receptors are present in the slime mold

Dictyostelium discoideum and the sponge Geodia cydonium [4]. The first structural signatures of rhodopsin-like GPCRs were found in several protostome Bilateria such as insects, molluscs, nematodes and trematodes, which indicates that rhodopsin-like receptors have an evolutionary age of 'only' 580–800 million years (My) [5-7]. Functional 5-hydroxytryptamine (5-HT) receptors seem to be among the oldest members of the rhodopsin-like GPCR family, as indicated by their presence in planarians [8], many nematodes [9] and vertebrates. Such a long evolutionary history emphasizes the importance of mining this source of structural information. However, the number and spectrum of GPCRs differ significantly between classes of species [10]. Vertebrates appeared 500 My ago (Mya), during the Cambrian period, and the number of their rhodopsin-like GPCRs is almost double that of invertebrates (excluding the odorant receptors). The total number of rhodopsin-like GPCRs is significantly smaller in more-basal vertebrate lineages and apparently increased during vertebrate evolution. This is predominantly because of a continuous accumulation of odorant receptor genes [11]. Intrachromosomal gene duplications [12], entire chromosomal [13] and whole genome duplications mechanistically underlie the gain of genes encoding GPCRs. In view of above a machine learning model has been developed to know the evolutionary origin of HIV by taking amino acid composition of bacteria, aves,other vertebrates, fungi and HIV [14-16].

Methodology:

To achieve our goal and develop our methodology we obtained the dataset from Swissprot/Uniprot databank of Expasy server [17]. The following two data sets were used.

Dataset1: It consisted of all the proteins of members of bacteria group. It consisted of 6 positive instances and other negative instances. All the entries marked as fragments were not included in the dataset.

Dataset2: It consisted of all the proteins of members of fungi group. All the entries marked as fragments were not included in the dataset. The total instances were 14 for fungi as positive instances and others are negative instances.

Dataset3: It consisted of all the proteins of members of animal group. The total instances of animal are 800 were taken as positive. All the entries marked as fragments were not included in the dataset.

Dataset4: It consisted of all the proteins of members HIV group. All the entries marked as fragments were not included in the dataset. The total instances of HIV were 50.

For training dataset we consider sequences belonging to all the datasets mentioned above. Support vector machine (Binary classification) is used for classification.

Support vector machine (Binary classification)

SVM is a supervised machine learning method which is based on the statistical learning theory [18.19]. When used as a binary classifier, an SVM will construct a hyperplane, which acts as the decision surface between the two classes. This is achieved by maximizing the margin of separation between the hyperplane and those points nearest to it. The SVMs were implemented using freely downloadable software, libSVM [20]. In this software there is a facility to define parameters and choose among various inbuilt kernels. They can be radial basis function (RBF) or a polynomial kernel (of given degree), linear, sigmoid.

.SVM software; LIBSVM

Simulations were performed using LIBSVM version 2.89 (a freely available software package) (20). For our study RBF Kernel was found to be the best. The SVM training was carried out by the optimization of the value of the regularization parameter and the value of RBF kernel parameter.

Amino Acid Composition

Previously, this parameter has been used for predicting the subcellular localization of proteins (21). The amino acid composition is the fraction of each amino acid type within a protein.

The fractions of all 20 natural amino acids were calculated by using Equation 1,

$$= \frac{Total\ Number\ of\ amino\ acid\ i}{Total\ number\ of\ amino\ acids\ in\ a\ protein}$$

Polycomp

The input vector of 450 was generated directly in the format of SVM by software Polycomp developed under Department of Bioinformatics, MANIT, Bhopal, India [22]. This software generates data which can be directly fed into the classifier hence saving valuable time and energy needed for formatting the hybrid.

Evaluation of Performance

The performance of our classifier was judged by 5 fold cross validation. The LibSVM provides a parameter selection tool using the RBF kernel: cross validation via grid search. A grid search was performed on C and Gamma using an inbuilt module of libSVM tools on Dataset 1, Dataset 2, Dataset 3, Dataset 4 as shown in Figure1, Figure2, Figure3, Figure4. Here pairs of C and Gamma are tried and the one with the best cross validation accuracy. On using the values of C=8 and Gamma=8.5 obtained through grid search an accuracy of 99.8626% was obtained on Dataset 1,the value of C=32 and Gamma= 0.5 obtained through grid search on Dataset 2,the value of C= 32.0 and Gamma= 0.125 obtained through grid search on Dataset 3,the value of C=0.5 and Gamma=8 are obtained through grid search on Dataset4.Prediction system assessment True positive (TP) and true negatives (TN) were identified as the positive and negative samples, respectively. False positives (FP) were negative samples identified as positive. False negatives (FN) were positive samples identified as negative. The prediction performance was tested with sensitivity (TP/ (TP+FN)), specificity (TN/ (TN+FP)), and overall accuracy (Q2). The accuracy for each group of HIV-1 was calculated as described by Hua and Sun [14] and shown below in equation 2.

$$Accuracy(x) = \frac{tp + tn}{tp + tn + fp + fn}$$

The graphs obtained are shown as:

Figure1 Grid search for bacteria GPCR dataset.

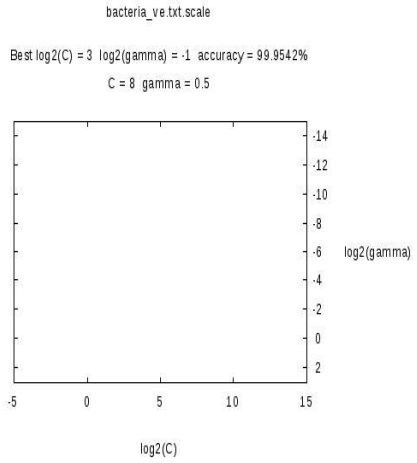


Figure 2: Grid Search for Fungi GPCR dataset.

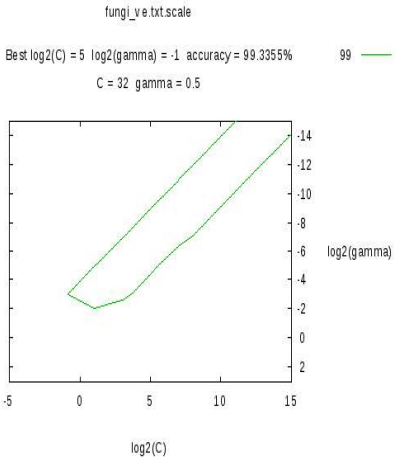


Figure 3: Grid search for Animal GPCR dataset.

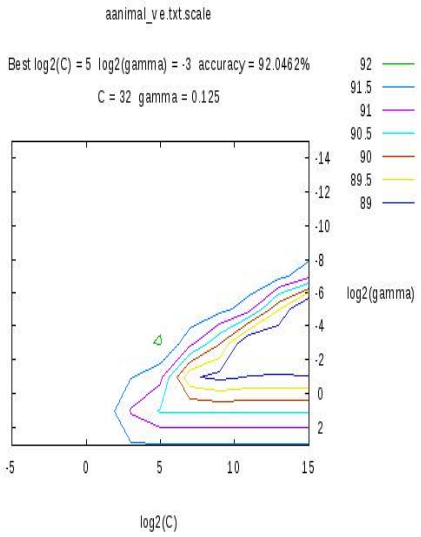
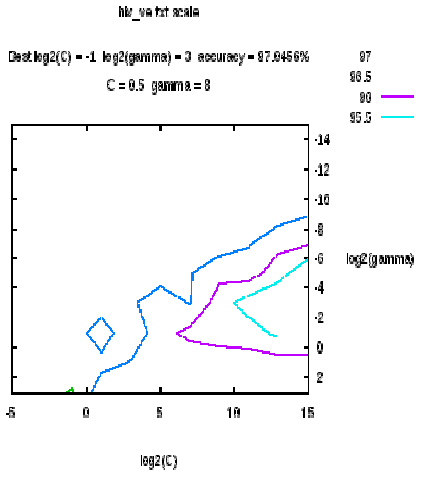


Figure 4: Grid search for HIV GPCR dataset.



Result & Discussion: Table 1. Shows the c, g, accuracy, error, recall, precision of groups bacteria, fungi, animal and HIV.

Group	c	g	Accuracy	Error	Recall	Precision
Bacteria	8	8.5	99.8626%	8.23%	0.00	0.00
Fungi	32	0.5	99.0698%	1.93%	0.00	0.00
Animal	32.0	0.125	99.6613%	15.78%	91.63	91.24
HIV	0.5	8	96.60%	3.04%	0	0

A “grid-search" on c and gamma (g) using cross-validation is performed. Various pairs of c and gamma (g) values are tried and the one with the best cross-validation accuracy is picked. It was found that trying exponentially growing sequences of c and gamma is a practical method to identify good parameters. The grid-search is straightforward but seems naive. In fact, there are several advanced methods which can save computational cost by, for example, approximating the cross-validation rate. However, there are two motivations why the simple grid-search approach is preferred. The grid-search can be easily parallelized because each (c; gamma) is independent. Since doing a complete grid-search may still be time-consuming using a coarse grid first. After identifying a better region on the grid, a finer grid search on that region can be conducted. Hence the grid search graphs for bacteria, fungi, HIV, animals have been plotted as shown in **Figures (1-4)**. As grid search depicts the performance, we can refer from the c, g, accuracy, error, recall, precision of groups bacteria,fungi,animal and HIV are given in Table 1. Precision and recall are two widely used metrics for evaluating the correctness of a pattern recognition algorithm. They can be seen as extended versions of accuracy, a simple metric that computes the fraction of instances for which the correct result is returned.The accuracies of bacteria, fungi, animal and HIV are 99.8626%.99.0698%, 96.60%.

The results obtained here will be helpful in differentiating between different groups and shows GPCR classification among bacteria, fungi, animal and HIV. A new GPCR protein discovered can be shown to either belonging to groups of proteins of bacteria, fungi, animals and HIV.Detecting functional regulatory elements within GPCRs but also for identifying signatures of past and ongoing evolutionary selection. The significance and reliability of results from such analyses depends on both the quality and the quantity of sequence data

obtained from different species and populations. In summary, harvesting ancient and modern sequence data will continue to be a powerful tool for both charting the history of evolution and obtaining information about GPCR structure and function. It will also help in the drug discovery and delivery.

V Acknowledgment

The authors are highly thankful to the Department of Biotechnology, New Delhi, India and M.P. Council of Science and Technology M.P., Bhopal, India for providing support in the form of Bioinformatics infrastructure facility to carry out the research work.

REFERENCES

[1] International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431, 931–945 2

[2]Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87

[3] Dobzhansky, T. (1973) Nothing in biology makes sense except in the light of evolution. Am. Biol. Teach. 35, 125–129

[4] Pin, J.P. et al. (2003) Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. Pharmacol. Ther. 98, 325–354

[5] Benton, M.J. and Ayala, F.J. (2003) Dating the tree of life. Science 300, 1698–1700

[6] Feng, D.F. et al. (1997) Determining divergence times with a protein clock: update and re-evaluation. Proc. Natl. Acad. Sci. U. S. A. 94, 13028–13033

[7] Peterson, K.J. and Butterfield, N.J. (2005) Origin of the Eumetazoa: testing ecological predictions of molecular clocks against the Proterozoic fossil record. Proc. Natl. Acad. Sci. U. S. A. 102, 9547–9552

[8] Saitoh, O. et al. (1997) Molecular identification of a G protein-coupled receptor family which is expressed in planarians. Gene 195, 55–61

[9] Tierney, A.J. (2001) Structure and function of invertebrate 5-HT receptors: a review. Comp. Biochem. Physiol. A Mol. Integr. Physiol. 128, 791–804

[10]. Fredriksson, R. and Schioth, H.B. (2005) The repertoire of G-proteincoupled receptors in fully sequenced genomes. Mol. Pharmacol. 67, 1414–1425

[11]. Mombaerts, P. (2001) The human repertoire of odorant receptor genes and pseudogenes. Annu. Rev. Genomics Hum. Genet. 2, 493–510

[12]. Perez, D.M. (2003) The evolutionarily triumphant G-protein-coupled receptor. Mol. Pharmacol. 63, 1202–1205

[13]. Ludwig, A. et al. (2001) Genome duplication events and functional reduction of ploidy levels in sturgeon (Acipenser, Huso and Scaphirhynchus). Genetics 158, 1203–1215

[14]. C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines. Software," available at http://www.csie.ntu.edu.tw/~cjlin/libSVM, 2001.

[15] A.Dubey, B.Pant and Neeru Adlakha,"SVM Model for Amino Acid Composition based Classification of HIV1 Groups". IEEE digital library published.

[16] A.Dubey, B.Pant and Usha Chouhan," SVM Model for Classification of Structural and Regulatory Proteins of HIV1 and HIV2 "Journal of Advanced Bioinformatics Applications and Research ISSN 0976-2604 Vol 2, Issue 1, 2011, pp 84-88

[17].www.uniprot.org

[18]C.C.Chang and C.J. Lin, "LIBSVM: a library for support vector machines, software, available at http://www.csie.ntu.edu.tw/~cjlin/libSVM, 2001.

[19] V. Vapnik, "The nature of statistical learning theory," Springer, 1995.

"Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition," Protein Eng. Des. Sel, vol. 17, pp. 509–516, 2004.

[21].Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, "Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect," J. Cell. Biochem, vol. 84, pp. 343–348, 2002.

[22]. Protein composition server, Department of Bioinformatics, MANIT, Bhopal, India, http://manit.ac.in/Procos.