Prediction and Classification for GPCR Sequences Based on Ligand Specific Features

Bekir Ergüner, Özgün Erdoğan, and Uğur Sezerman

Biological Sciences and Bioengineering, Sabanci University, Orhanli – Tuzla
34956 Istanbul, Turkey
{bekir, ozgune}@su.sabanciuniv.edu,
ugur@sabanciuniv.edu

Abstract. Functional identification of G-Protein Coupled Receptors (GPCRs) is one of the current focus areas of pharmaceutical research. Although thousands of GPCR sequences are known, many of them are orphan sequences (the activating ligand is unknown). Therefore, classification methods for automated characterization of orphan GPCRs are imperative. In this study, for predicting Level 1 subfamilies of GPCRs, a novel method for obtaining class specific features, based on the existence of activating ligand specific patterns, has been developed and utilized for a majority voting classification. Exploiting the fact that there is a non-promiscuous relationship between the specific binding of GPCRs into their ligands and their functional classification, our method classifies Level 1 subfamilies of GPCRs with a high predictive accuracy between 99% and 87% in a three-fold cross validation test. The method also tells us which motifs are significant for class determination which has important design implications. The presented machine learning approach, bridges the gulf between the excess amount of GPCR sequence data and their poor functional characterization.

Keywords: G-Protein Coupled Receptors (GPCRs), ligand specificity, GPCR sequence.

1 Introduction

GPCR (G-Protein Coupled Receptors) are a large family of trans-membrane proteins responsible for signal transduction. The GPCRs receive various external stimuli ranging from chemical to physical and in turn activate intracellular G-proteins. Cells can accept and respond different extracellular and physical signals. Acceptance of a signal occurs principally in two different transduction pathways. One is mediated by tyrosine kinase receptors and the other by G protein-coupled receptors (GPCR)[4].

GPCR is an essential subject of many recent biomolecular projects. They are responsible for diverse physiological processes such as neurotransmission, secretion, cellular metabolism rowth and cellular differentiation as well as inflammatory and immune responses. Therefore, they are vital for the research and development for new drugs [3].

A. Levi et al. (Eds.): ISCIS 2006, LNCS 4263, pp. 174 – 181, 2006.

Various databases have been created in order to observe and categorize different characteristics of GPCRs. These databases hold sequences, mutation data and ligand binding data. Moreover, these databases are further improved by multiple sequence alignments, two dimensional visualization tools, three dimensional models and phylogenetic trees [5].

Even though thousands of GPCR sequences are known as a result of ongoing genomics projects , the crystal structure has been solved only for one GPCR sequence using electron diffraction at medium resolution (2.8 A) to date and for many of the GPCRs the activating ligand is unknown, which are called orphan GPCRs [11]. Hence, based on sequence information, a functional classification method of those orphan GPCRs and new upcoming GPCR sequences is crucial to identify and characterize novel GPCRs.

In the current literature, to classify GPCRs in different levels of families, there exist different attempts, such as using prim database search tools, e.g., BLAST [1], FASTA [8]. However, these methods only work if the query protein sequence is highly similar to the existing database sequences in order to work properly. In addition to these database search tools, the same problem is addressed by using Hidden Markov Models [10], bagging classification trees [6] and SVMs [7]. One other method studies the tertiary structure of GPCRs by using only the amino acid sequence (MembStruck) and the binding site and binding energy of various ligands to GPCRs (HierDock) [9]. Out of all these methods Karchin et al. (2001) showed that SVMs gave the highest accuracy in recognizing GPCR families [7]. In SVMs, an initial step to transform each protein sequence into a fixed-length vector is required and the predictive accuracy of SVMs significantly depends on this particular fixed-length vector. Karchin et al., pointed out that the SVM performance could be increased by using most relevant feature vectors, since SVMs do not identify the features most responsible for class discrimination. Therefore, for an accurate SVM classification, feature vectors should reflect the unique biological information contained in sequences, which is specific to the type of classification problem. In a recent work Bakir et al. used a fixed length feature vector of 40 most distinguishing patterns to classify amine sub-family GPCRs with 97% accuracy using SVMs[2].

2 System and Methods

In this paper we used several machine learning approaches to classify GPCRs according to their ligand specificities rather than their subfamilies. The GPCR groups we chose to work on were selected according to the ligands they bind to: amines, peptides, olfactory and rhodopsin. The binding of ligand to GPCR occurs outside of the cell therefore to understand interaction of GPCR and ligands; we decided to examine the primary sequence information of extracellular regions of GPCRs. Since The GPCR is a 7TM protein, meaning that it has 7 trans-membrane regions, the regions we observed were an N-terminus and three extracellular loops. We acquired the GPCR amino acid sequences from GPCRDB database which also groups GPCR proteins into subfamilies (http://www.gpcr.org/7tm/multali/multali.html) [12]. Total of 352, 1998, 595, 355 and 56 proteins from *amine*, *olfactory*, *peptide, rhodopsin* and *prostanoid* subfamilies derived from GPCRDB respectively. After derivation of

proteins from database their secondary structure is determined by using TMHMM (trans-membrane hidden Markov model) server. The dataset is available upon request. Thus we could isolate the n-terminus and extracellular loop sequences. Using different alphabetic coding systems for amino acids, we created a database consisting of amino acid triplet frequencies of each extracellular sequences for each ligand class studied.

Our database was created using MYSQL due to its user friendly nature. Another important factor in choosing MYSQL as our database system was being able to control the MYSQL database through Microsoft Visual C++ by MYSQL++ implementation (http://tangentsoft.net/mysql++).

We randomly separated each existing GPCR group into 2 subgroups as 'train' and 'test' in a 2:1 ratio for amine, rhodopsin, prostanoid subfamilies, 5:1 ratio for peptide subfamily and 9:1 ratio for olfactory subfamily. For both subgroups, the amino acid sequence of n-terminus, loop1, loop2 and loop3 regions were grouped into triplets. For 'n' amino acid long sequences, this would provide us with 'n-2' possible triplets. Using triplets seemed to be the optimal choice since using single amino acids would not help to determine neighborhood information in the sequence. The reasoning behind this was to focus on more specific patterns in the amino acid sequence while not losing vital patterns. Since it is too specific, using 5-amino-acid bundles would greatly diminish the number of matches, possibly ignoring positive matches that would have been spotted using triplets.

The 'train' subgroups were loaded into the database to create the basis for prediction patterns which would later be applied to the 'test' subgroup. These final results would show us how efficient the prediction patterns were.

The results of these pattern searches provided us with certain information such as the number of triplet occurrences and in how many proteins a certain pattern was spotted. In order to find specific patterns for different ligand groups we compared the results of each group with one another. The over-expression of a specific pattern in one of the two GPCR groups compared showed us that this pattern was characteristic of that certain GPCR group. However, this method was not precise enough since a pattern that separates amines-peptides might not do so for amines-olfactory. Therefore, to further enhance our statistical precision, we compared the results of each group to that of all the remaining GPCR groups combined.

We also applied an index search on amino acid sequences to check whether the positions of dominant triplets carried an importance in the separation of GPCR groups. We were hoping to spot a parallelism between triplet densities and their positions in the extra cellular portions of the protein.

All of these comparison methods were first used on a 11-letter amino acid alphabet (Table 1), then repeated on 6-letter and 20-letter alphabets. The usage of different alphabet systems allowed us to examine the effect of a single amino acid compared to the general biochemical properties of the group it belongs to. The classifications of amino acids were done according to their physical and chemical properties. Evolution allows for conserved mutations that do not change the physical and chemical nature of mutation site since these mutations do not disrupt the function of the molecule. By

Table 1. Classification of amino acids

Class	Amino Acid(s)	Class	Amino Acid(s)
a	I,V,L,M	g	G
b	R,K,H	h	W
c	D,E	i	C
d	Q,N	j	Y,F
e	S,T	k	X,P
f	A		

using classification of amino acids we can capture this nature of evolution. Since the patterns we expect to observe can vary allowing acceptable mutations.

For each comparison of classes we found the ratio of existence of a triplet at a certain location (e.g loop 1) in one class against the other class. We ranked the triplets with these ratios. The words with highest ratios would make up the important features. This is done for six times comparing each class to another class. These important features are selected from training set only. Then starting from the highest ratio we search how many sequences in the training set can be selected with this motif only. Going down the ratio ranking we add a new motif to important motif list if it helps to identify new sequences other than the previous motifs. We carry on until no new sequence can be identified with the rest of the motifs. This is done for each classification problem and we get a list of important motifs for each class. Then given the test sequence we look at existence of these motifs in the query sequence. We assign the query sequence to the class that has the highest hits on the query sequence. Unlike the patterns obtained in Bakir's work, the selected patterns are observed only in the specific group and not the other groups [2]. In the previous work, selected patterns had to be present in at least 50% of the sequences in the selected group, presence of the same pattern in other classes were not checked.

3 Results

The motifs obtained from amine peptide comparison are listed in Table 2. Showing how many sequences in the training set they occur and how many new sequences they help to identify. We keep the motifs that cause no misclassification between these classes. In determining the most important motifs the patterns that occur in most of the given class that do not occur in the other classes are ranked according to number of occurrences in the given class. The most important motif distinguishing amine from the peptide is existence of word bhe in loop1. Since it helps to classify 33 amine GPCRs without any misclassification.

By running all possible comparisons of classes we obtain the list of important features that helps us to identify a given class. The list is derived from ranked

Table 2. Occurences of motifs for amine versus peptide classification

n-terminal			**loop2**			**loop1**			**loop3**		
motif	occur	new	motif	occur	New	motif	occur	new	motif	Occur	new
Ddh	13	13	dkf	26	26	bhe	33	33	ibc	27	27
Ief	11	11	cij	22	21	egb	27	12	bak	20	19
Dif	10	3	ckg	19	14	aic	26	23	fai	19	12
Fkh	8	8	kbi	15	12	hka	25	9	kia	13	13
Hfa	7	3	dic	14	14	ihj	16	16	abk	12	8
Chc	6	6	ggi	12	12	afi	15	14	kkc	12	6
Hff	6	5	kdi	12	8	ejg	14	4	cid	10	9
Ffh	5	5	iic	11	3	abi	13	13	jbk	9	4
Hck	5	3	bij	10	5	aha	12	7	kcj	9	3
Hdg	5	5	cci	10	9	bah	12	8	gke	8	8
Fbh	4	3	jfa	10	4	jhf	12	11	jid	7	6
Dhk	3	3	eki	5	4	ehi	11	9	jef	5	4
			gdi	5	5	fgj	10	10	dai	4	4
			hdd	4	3	jic	8	3	ibf	4	3
			jbi	4	4	bfh	6	6	aie	3	3
			cig	3	3	bhc	4	4	bfh	3	3
						fjh	3	3	icc	3	3
									jig	3	3

important features table. We start from the most distinguishing feature and add on a new feature if it helps us to classify a new GPCR sequence correctly. The motif list ends when the remaining motifs s can only classify already distinguished sequences thus yielding to no new classifications. Using these motif lists, we try to determine the class of the sequences in the test set. The results can be seen in Table 3. Success rate of classification varies between 87% and 99% and we can also determine the important motifs for each class.

We had problem with prostanoids therefore we eliminated them from our search. There were only few prostanoids to be able determine any kind of significant patterns. The search yielded very few patterns with such stringent determination of patterns. Therefore any class compared with the prostanoid binding group gave more hits. Therefore only 28% of the prostanoids could be identified and all the others were classified to other classes. Currently we are allowing for more errors in pattern identification to overcome this problem.

In order to check how accurate the results were, we used a program called CART, a software program or building regression trees. For testing the accuracy of patterns that recognize amines, 40 patterns were given as predictors and 18 test proteins were used. The result of patterns comparing amine versus all ligand classes are shown in Figure 1. The importance of the patterns in classification of amines are summarized in Table 4. Unlike the patterns used in the previous method, the patterns used in CART distinguish one ligand class from all the other 4 ligand classes. For example distinguishing patterns for amines used in CART are the best patterns in the combination of amines vs. peptides, amines vs. olfactory, amines vs. rhodopsin and amines vs. prostanoid pattern sets in the first method. Thus patterns in Table 4 can differ from the ones in Table 2. For example pattern ‘caa’ is the most important pattern in Table 4 but it is not seen in Table 2. This is because ‘caa’ is important for amines vs. rhodopsin and amines vs. prostanoid classification but not as important for amines vs. peptides and amines vs. olfactory classification. Since Table 2 shows the patterns for amines vs. peptides, ‘caa’ pattern is not seen.

The main novelty of this method is to determine motifs using reduced alphabet representation and using information theory for determining significance of the motifs. This increased the prediction accuracy drastically while enabling the end users (pharmaceutical companies) to determine significant motifs for ligand determination that can be used for drug design purposes.

Table 3. Total success rates of classifications

	Correct	**total**	**success**
amines	104	120	0.87
peptides	98	102	0.96
olfactory	195	195	1
rhodopsin	87	111	0.87
prostanoid	5	18	0.28

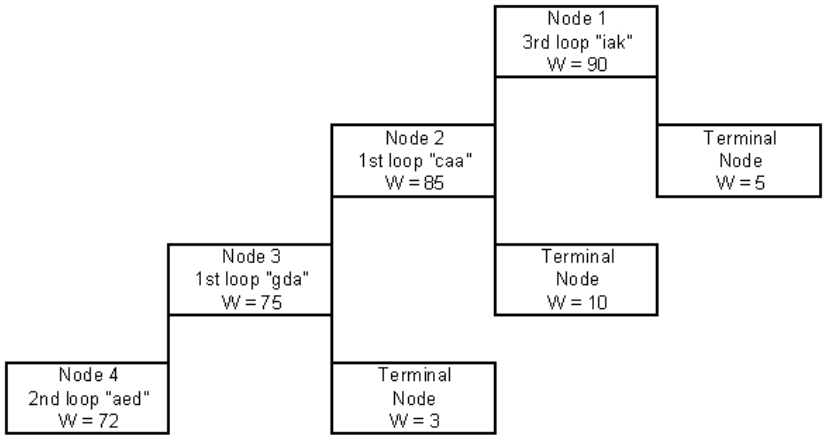

Fig. 1. The classification table showing the only patterns determining amines from all others. The figure shows that pattern "iak" occurring in 3rd loop of extracellular region is the most crucial pattern determining amines of all ligand groups with weight 90. The increasing number of the nodes are in the decreasing order of importance of determining amines. W is the number of sequences populating a given node.

Table 4. Variable importance of the amine determining patterns

Patterns	Relative Importance
Loop 1 'caa'	100
Loop 1 'gbh'	97.46
Loop 3 'iak'	83.767
Loop 1 'gjh'	64.62
Loop 1 'gda'	51.101
Loop 2 'aed'	44.942
Loop 1 'agj'	43.636
Loop 1 'aag'	31.099
Loop 1 'dca'	22.736
Loop 3 'akc'	17.737
Loop 1 'hjj'	16.511
N-term 'afa'	12.811
N-term 'eea'	0

Acknowledgements

The authors wish to thank to Alper Kucukural, Gurkan Yardimci, Berkay Kaya, Hanife Kebapci, Emir Tinaztepe, Nalan Liv and Yekta Yamaner for their help in different parts of the project.

References

1. Altshul,S. et al. Basic local alignment search tool. *J. Mol. Biol.*, 215, (1990) 403-410
2. Bakir ,B. Sezerman, U.Functional Classification of G proteins based on their specific ligand coupling patterns, LNCS, 2006
3. Bouvier, M. Structural and functional aspecs of g protein-coupled receptor oligomerization. *Biochem. Cell Bio.* 76, (1998) 1-11
4. Thomas Gudermann, Torsten Schöneberg, and Günter Schultz. Functional And Structural Complexity of Signal Transduction via G-protein Coupled Receptors. *Annu. Rev. Neurosci.* 20, (1997) 339-427
5. F. Horn, J. Weare, M. W. Beukers, S. Hörsch, A. Bairoch, W. Chen, Ø. Edvardsen, F. Campagne and G. Vriend, GPCRDB: an Information System for G Protein-Coupled Receptors. *Nucleic Acids Res.* 1, (2003) 294-7
6. Huang,Y. et al. Classifying G-protein Coupled receptors with bagging classification tree. *Computational Biology and Chemistry,* 28, (2004) 275-280
7. R. Karchin, K. Karplus, D. Haussler. Classifying G-protein Coupled Receptors with Support Vector Machines. *Bioinformatics.* 18, (2002) 147-59
8. Pearson,W. and Lipman,D. Improved tools for biological sequence analysis. *Proceedings of National Academic Science*, 85, (1988) 2444–2448. Database search tool is available at: http://www.ebi.ac.uk/fasta33
9. N. Vaidehi, W. B. Floriano, R. Trabanino, S. E. Hall, P. Freddolino, E. J. Choi, G. Zamanakos, W. A. Goddard III. (2002) *PNAS*. 99,20,12622-7
10. Sreekumar,K.R. et al. Predicting GPCR-G-Protein coupling using hidden Markov models, *Bioinformatics,* 20, (2004) 3490-3499
11. Tusnády,G.E. and Simon,I. The HMMTOP transmembrane topology prediction server. *Bioinformatics*, 17 (2001) 849-850. Available at: http://www.enzim.hu/hmmtop
12. G Protein-Coupled Receptor Data Base, http://www.gpcr.org/7tm/multali/multali.html