A SVM for GPCR Protein Prediction Using Pattern Discovery

Francisco Nascimento Junior, Ing Ren Tsang and George Darmiton da Cunha Cavalcanti
Center of Informatics (CIn), Federal University of Pernambuco (UFPE),
P.O. Box 7851, Cidade Universitária,
Cep: 50.740-530 - Recife - PE - Brazil
(e-mail:fnj,tir,gccd@cin.ufpe.br)

Abstract

Machines learning techniques have been applied in several different problems in bioinformatics. Similarly, pattern discovery algorithms have also been used to uncover hidden motifs in protein sequences, contributing greatly to the understanding of the problem of protein classification. G-protein coupled receptors (GPCRs) represent one of the largest protein families in Human Genome. Most of these receptors are major target for drug discovery and development. Therefore, they are of interest to the pharmaceutical industry. The technique used in this paper combine machine learning and pattern discovery methods to develop a protein prediction procedure in relation to its functional class, more specifically to predict GPCR protein class. Vilo[2] proposed an algorithm in order to extract pattern of regular expressions from known protein GPCR sequences and used them to predict coupling specificity of G protein coupled receptors to their G proteins. We analyze these patterns and combine them as features for feeding a SVM to predict the GPCR super class. We demonstrate the results using ROC curves, which are well-indicated to evaluate the performance of this kind of classifiers. The experiments, based on the GPCRDB database, also showed that we were able to find some novel GPCR sequences that were not described in the PROSITE database.

Keywords: SVM, GPCR, Pattern Discovery, Prediction, Proteins

1 Introduction

G protein-coupled receptors (GPCRs), is also known as seven transmembrane receptors or 7TM receptors because of its structural characteristic. They represent one of the largest protein families in the human genome. Also, they are major target for drug discovery and development because approximately 50% of these receptors appear to be of great relevance to the pharmaceutical industry and 40% to 60% of the current drugs on the market, target GPCRs. A very important aspect of their function is the coupling specificity with members of G-proteins families. A GPCR can interact with one or more G-proteins; an interesting problem is the prediction of the coupling specificity of GPCRs to the G-protein family class. Several prediction methods have been developed to accomplish successfully this task [1, 4, 5, 6, 7, 13, 14, 15, 16]. However, we are here interested in solving a different problem, which is the prediction of a GPCR protein from an unknown class of protein, instead of its coupling specificity. This procedure is of interest to identify GPCR proteins that are not yet annotated.

Based on the patterns of regular expressions found by Moller et al. [1], we used a learning procedure to predict if a protein belongs to a GPCR protein class or not. Like some other authors [13, 14, 15, 16], the main objective of Moller's el al. work was to device a method to predict the coupling specificity class of a GPCR protein to be of either Gi/o, Gs or Gq/11 class. Diferently, here we would like to develop a method to predict if an unknown protein sequence belongs to the GPCR protein class or not. This problem is of interest since the discovery of new GPCR proteins is of great interest for the pharmaceutical and biotech industry.

A class of statistical learning algorithms called Support Vector Machines (SVMs) presented by Vapnik became quite popular in the machine-learning community during the 1990s [11]. When SVMs are applied to the simplest learning problem, two-class pattern recognition, the learning machine shows a series of labeled examples from two categories and is trained to distinguish between them. Karchin et al. developed successfully a classifying system for G-protein coupled receptor using a SVM [13].

2 Methods

In article [1], a set of 40 patterns of regular expression were obtained to characterize the coupling receptors for each of the three coupling classes. These patterns were ac-

quired by using the SPEXS patterns search program [2]. A classification system was constructed to successfully predict the GPCR coupling classes based on derived pairs and triplets of those patterns. The classification was obtained by simple measurements of specificity and sensitivity of the patterns to the best class classification.

The success of Moller's et al. classification scheme based on 120 patterns and the high values of the specificity of each pattern suggest that these patterns are appropriate for not just detecting coupling specificity classes of GPCR proteins but it also can be used to characterize a general GPCR protein from another different type of protein. Using the 120 patterns as features for GPCR protein prediction method, we have made an implicit assumption that the patterns well qualify a GPCR protein. Even though in [1] the original authors constrained the pattern to the regions that were candidates for the physical interaction with the G-protein, the intracellular loops and the C-terminus, our approach is to use the total number of patterns as signature for a classifier. We first test our assumption using a simple chi-square statistics, to evaluate how well the patterns will perform in a first classification attempt, and secondly we use a more elaborate classifier system namely SVM based on statistical learning.

Table 1, 2, and 3 of the original article [1] shows the patterns of regular expressions that were used to classify the G-protein family, each column represents the specific coupling that the patterns best represents.

2.1 Feature extraction characteristics

Extracted characteristics matrix were be used as input for the classifier. This matrix contain values obtained from running String Matching algorithm in the patterns from Vilo and the sequences, thus the number of occurrences for each pattern per sequence represented by a matrix n x p, where n is the number of sequences and p, the number of patterns. Based on these values, we created three types of matrices:

1. Type A: represents only the number of occurrences of patterns per sequence.

2. Type B: represents the sum of the length of the substrings matched of patterns per sequence.

3. Type C: represents the position of the occurrence of substring matched of patterns per sequence.

2.2 Chi-square test

Chi-Square test is a statistical procedure used to evaluate the goodness of the fit from a known distribution compared to another distribution [8]. The equation is defined as:

$$\chi^2 = \sum_{i=1}^{p} \frac{(M_i - m_i)^2}{m_i} \quad (1)$$

where, Mi represents the average value of the curve we want to compare and mi is equal to the average value obtained from the known data. The index "i" represents all 120 different patterns. This means that the smaller the value of the test, the closer the curve is to the known distribution. We will use two distributions: one to characterize the GPCR protein and another to characterize all the others protein type (or non-GPCR protein type), so that we can compare the distribution of the sequence that we are interested in predicting to both known distribution that characterize either GPCR or non-GPCR protein.

The main objective to use the Chi-Square test is to verify if the frequency of a specific observed event in a sample has strayed or not significantly from expected frequency. In our case, we would want to consider if the analysis of frequency of matches of pattern set will differentiate the sequences of GPCR class against sequences of Non-GPCR class.

2.3 Support Vector Machine

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers. They can also be considered a special case of Tikhonov regularization. A special property of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin; hence they are also known as maximum margin classifiers.

Support vector machines map input vectors to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data. The separating hyperplane is the hyperplane that maximizes the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be [11].

2.4 Hypothesis testing

Two hypotheses were tested here, first if the proposed patterns of regular expressions are indeed suited for classifying the different GPCR coupling receptors using a SVM. The second hypothesis is a generalization of the first, in respect to the characterization of GPCR proteins type through the use of these patterns into respect of others protein types. In short, if these patterns are able to separate GPCR from non-GPCR proteins.

Moller et al. used a classifying system which takes into account the occurrence of combined pairs and triplets of

patterns in the sequence. Since they tested the procedure in protein sequences that they knew were GPCR, they could isolate the transmembrane region and use statistics of the occurrence of the pattern to predict the coupling specificity of the G protein coupled receptors to their G proteins. Since we are interested in the prediction of a GPCR protein as a sequence we will use the whole sequence length to obtain the patterns features of our classifying system.

In order to verify if the patterns obtained using the whole protein sequence have similar statistical characteristics as the ones obtained using just the trasmembrane region, we analyze the pair of pattern occurrence in its different coupling domain.

Since it would be extremely cumbersome process to combine all patterns in tuplas. We use a machine learning approach which should test all possible combinations and uses those that gives the better results for classification.

2.5 GPCR and non-GPCR

Our main objective is to answer the question: is an unknown sequence a GPCR protein? So we analyzed the pattern of co-occurrence from data of proteins sequences that we know to be GPCR, in contrast to sequences that we know that belongs to proteins that are non-GPCRs. A difference between total number of patterns for GPCR and non-GPCR sequences emerged from this analysis. Here we want to test the possibility to use these patterns to derive a distinction from the class function protein sequence. An analysis for a group of GPCR's and a group of non-GPCR's is presented.

2.6 ROC

A receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot of the sensitivity versus (1 - specificity) for a binary classifier system as its discrimination threshold is varied. ROC can also be represented by plotting the fraction of true positives (TPR = true positive rate) vs. the fraction of false positives (FPR = false positive rate). Also known as a Relative Operating Characteristic curve, because it is a comparison of two operating characteristics (TPR & FPR) as the criterion changes [12], the ROC curve shows the ability of the classifier to rank the positive instances relative to the negative instances. An ROC curve is a two-dimensional depiction of classifier performance. To compare classifiers we may want to reduce ROC performance to a single scalar value representing expected performance. A common method is to calculate the area under the ROC curve, abbreviated AUC. Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1. So that the close the AUC value is to 1 the better is the performance of the system.

3 Experiments

In this experiment, it was used a set of 769 sequences of GPCR from GPCRDB and 2565 of non-GPCR, these are the same set used in [13]. Hence, it was created a matrix of characteristics which size 3334 x 120 containing for each type defined in section 2.1.

Distribution plot: Using a set of GPCR and Non-GPCR protein sequences, we generated a distribution plot based on the 120 patterns suggested above. These plots can provide some insights on the performance of these patterns as features to characterize GPCR from Non-GPCR sequences. In Figure 1, we show a histogram plot of number of occurrence into using only 30 patterns.

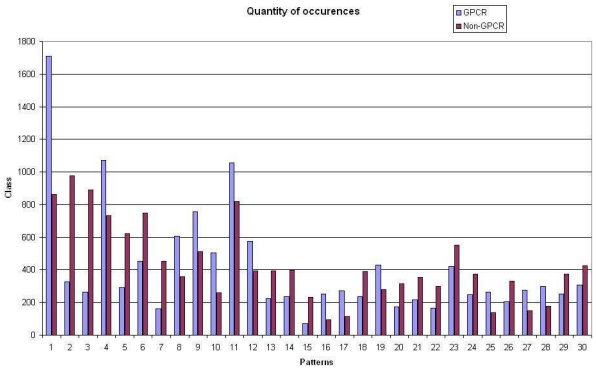

Figure 1. Histogram of pattern occurences

The difference in the distribution gives an idea of how a classifier can perform. This plot shows a significant difference between GPCR and non-GPCR. The next step is to apply a simple chi-test statistics to verify if this difference will be sufficient to make the recognition possible.

4 Results

4.1 Chi-Square test

Matrices generated using extraction Type A and Type C were used to analysis distribution of GPCR and non-GPCR sequence. The null hypothesis used was that the frequency of GPCR in the matrix is not different from the frequency of non-GPCR. Therefore, the alternative hypothesis defends that there is such a difference. Using a matrix of occurrences, it was calculated mean and standard deviation for both the subsets. Applying the chi-square equation shown in the previous section the following result was obtained:

The results implies that the null hypothesis was rejected in both types, concluding that frequencies of GPCR and Non-GPCR are, statistically different, as well one as other.

Table 1. Hypothesis Test from Samples of Type A and Type C

	Type A		Type B	
Samples	Mean	Std Dev.	Mean	Std Dev.
GPCR	219,00	224,94	32,6875	46,8634
Non-GPCR	208,90	211,49	13,6786	22,5581
X2	47,63		5,00	
Critic value = 3,841 (5% of significance)				
	Rejected		Rejected	

Table 2. Better results from experiments using MLP

Config	Type	Epochs	L. Rate	Hidden	AUC
1	B	4000	0.35	10	**0.87443**
2	B	1000	0.25	10	**0.87397**
3	A	4000	0.05	10	**0.86174**
4	B	4000	0.35	10	**0.84249**

So, we conclude it is possible to use such frequencies like characteristics for classifying. We decided don't test matrix of type B, because this type is derived through by type A, therefore, if type A was rejected, B also would be.

4.2 MLP

We decided to realize experiments using MLP to order to have results to compare with others. The used implementation was from Neural Network Toolbox of Matlab. The training was done with sigmoid logistic as transfer function between layers, backpropagation training algorithm and, for finding better setup of machine, we varied the parameters of epoch, learning rate and number of hidden neurons. In table 2, it is showed better AUC-values for these experiments.

4.3 SVM

A more elaborate way to predict the protein class was tested using SVM. This machine was chosen based on good results showed in [13]. For experiments, we used the same set of sequence, Matlab 7.0.4 and a implementation of SVM and ROC from [17]. For beginning, the training of the machine was done varying the parameters C (bound on the lagrangian multipliers) in the range from 2^{-6} to 2^{6}, λ (conditioning parameter for QP method) from 1^{-7} to 1^{-2}, kernel function gaussian or polynomial and a parameter of kernel which we called kernelOption, using values in $\{0.5, 1, 5\}$. Furthermore, it was used data generated by three cited types of extraction of characteristics. Combining all of these values and bases, it was trained the machine using crossvalidation of n-fold (n ranging from 2 to 10). To resume, we have 163 instances of experiments using four parameters: type of matrix (A, B, C), C-value (4, 8), OptionKernel(1,5,10) e n-fold (2 to 10). In table 3, we show some lines of these instances and we selected better results, using as criterio of order, the higher average of AUC-value obtained grouping by all of parameters. In table 3, we can see 5 highers values of AUC for the experiments using SVM.

Table 3. Better results from experiments using SVM

Config	Type	C	K. Opt	N-Fold	Avg AUC	Std AUC
1	A	4	10	9	**0,98961**	0,00641
2	A	4	10	10	**0,98934**	0,00644
3	A	8	10	9	**0,98903**	0,00655
4	B	4	10	9	**0,98884**	0,00700
5	A	8	10	10	**0,98882**	0,00692
6	A	4	10	6	0,98879	0,00699
7	A	8	10	6	0,98836	0,00684
8	A	4	10	8	0,98827	0,00771
9	B	8	10	9	0,98817	0,00724
10	A	4	10	3	0,98800	0,00596
11	B	4	10	6	0,98789	0,00586
12	A	4	10	7	0,98784	0,00796
13	A	8	10	8	0,98768	0,00762
14	A	8	10	7	0,98761	0,00817
15	A	8	10	3	0,98747	0,00609

4.4 One-class Classifier (OCC)

The classification based on just one class data is a special case of general problem of classification. In this case, we treat the problem in the same situation as with two classes classification, having each class a specific meaning, which can be called target and outlier class:

- target: this is the class that we assume is well represented, that is there are many examples for training. The training set is not necessarily completely composed according the distribution found on data set. The sample used in this target class can be obtained from the ideal representation of the target data.
- outliers: The problem of this class is that it can be very sparse or totally absent, very hard to measure, or yet, can be very expensive to generate it. In principle, a classifier of one-class would be able to work just with target examples. And, in extreme cases, it's possible also that the outliers be so abundant that a good sample of this data would not be possible.

4.4.1 Minimization of errors

To finding a good one-class classifier, two types of errors must be minimized: the false positives rate and the negatives falses rate. The advantage of this type of classifier in a problem such as protein prediction is that we would just need to know a data set that represents well the known protein, other then having to generate a data set of the unknown types. In our case we would just need to have a well curated database such as GPCRDB.

4.4.2 Experiments

We used ddtools, a Matlab toolbox for Pattern Recognition, to implement the one-class classifier and, we performed experiments with 10-fold cross-validation. With this toolbox, it's possible to use many algorithms as method of classification according the options below:

- (1) Gaussian
- (2) Mixture of Gaussians
- (3) Incremental SVM with linear kernel
- (4) Incremental SVM with polynomial kernel
- (5) Incremental SVM with exponencial kernel
- (6) Incremental SVM with kernel of radial basis

The obtained results are showed in table 4.

Table 4. Better results from experiments using OCC

Config	Type	Algoritm	Avg AUC	Std AUC
1	A	4	0,947638	0,03838
2	B	4	0,944142	0,03952
3	C	1	0,941438	0,03870
4	C	2	0,937505	0,03892
5	C	6	0,922445	0,06739
6	A	6	0,906255	0,07644
7	B	6	0,902734	0,08018
8	A	1	0,889690	0,06704
9	A	2	0,886820	0,08062
10	B	1	0,885910	0,06863
11	B	2	0,883470	0,08622
12	C	4	0,800860	0,12106
13	A	5	0,641330	0,18094
14	B	5	0,629630	0,18622
15	C	5	0,563390	0,19955
16	A	3	0,414530	0,20024
17	C	3	0,345310	0,18655
18	B	3	0,321300	0,20428

5 Conclusions and Discussion

A procedure for identifying possible GPCR protein based on patterns of regular expressions and classification scheme based on SVM was proposed. This procedure also allows verification of the coupling specificity of a GPCR. The basic classification of GPCR protein relies on the analysis of the number of hydrophobic regions that would be candidates for transmembrane regions. Any protein with seven such helices would be a prime candidate for a GPCR. Differently from this technique, the proposed method uses a general pattern search throughout the whole protein sequence in the process of identify the type of protein. We successfuly tested our method using the GPCRDB database, however we were also able to corrected identify nine not trained GPCR proteins sequences described in the GPCRDB but not described in the PROSITE database, thus showing that the method would be able to revel novel GPCR protein sequences.

Another interesting point is the ability of the method to assign the coupling specificity the GPCR. And differently from the original Moller et al. article the method proposed uses a SVM as a classifier which can also uncover multiple coupling. Protein functional prediction, in general, is a very challenger field, where no single technique has been proven to work without fail. Structural based methods are often used to improve the prediction power of the method. Even though several techniques from the literature apply just the primary sequence information, this certainly will prove limited if one is looking for a close to 100% prediction. The method described here will also be limited in scope. Since it is very general (in the context that it does not make use of (biological) information specific to the protein class), but on the other hand, it can be easily applied to classify different types of proteins. The constraint of the method lies basically only in the selection of known sequences, since we could use the same framework of pattern discovery and classification to any other type of protein, the only constraint would be to obtain a very well represented data set. Presently we are investigating this procedure using kinases proteins.

References

[1] Moller, S., Vilo, J. and Croning, D.R.: Prediction of coupling specificity of G protein coupled receptors to their G proteins. Bioinformatics,**17** (2001) S174–S181

[2] Vilo, J.: Discovering Frequent Patterns from Strings. Department of Computer Science, University of Helsinki (1998).

[3] Wang, J. T. L., Shapiro, B.A. and Shasha, D.: Pattern Discovery in Biomolecular Data: Tools, Tech-

niques, and Applications. Oxford University Press, USA. (1999).

[4] Cao,J., Panetta, R., Yue,S., Steyaert, A., Young-Bellido, M. and Ahmad, S.: A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. Bioinformatics, **19**, (2003) 234-240.

[5] Nikolaos G. Sgourakis, Pantelis G. Bagos and Stavros J. Hamodrakas: Prediction of the coupling specificity of GPCRs to four families of G-proteins using hidden Markov models and artificial neural networks. Bioinformatics, **21**, (2005) 4101-4106.

[6] Kodangattil R. Sreekumar, Youping Huang, Mark H. Pausch and Kamalakar Gulukota: Predicting GPCR-G-protein coupling using hidden Markov models. Bioinformatics, **20**, (2004) 3490-3499.

[7] Pierre Baldi and Soren Brunak: Bioinformatics: The Machine Learning Approach: The MIT Press, Cambrigde USA (1998).

[8] William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery: Numerical Recipies in C++: The Art of Scientific Computing, Cambridge University Press (2002).

[9] Simon Haykin: Neural Networks: A Comprehensive Foundation: Prentice Hall (1998).

[10] Martin Riedmiller and Heinrich Braun; A direct adaptive method for faster backpropagation learning: the Rprop algorithm: Proceedings of ICNN, San Francisco (1993)

[11] Christopher J.C. Burges; A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2:121 - 167, 1998.

[12] Tom Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers. HP Laboratories, March 16, 2004.

[13] Rachel Karchin, Kevin Karplus and David Haussler. Classifying G-protein coupled receptors with support vector machines. Bioinformatics. Vol. 18:147-159 (2002)

[14] Matthew N. Davies, Andrew Secker, Alex A. Freitas, Miguel Mendao, Jon Timmis and Darren R. Flower. On the hierarchical classification of G protein-coupled receptors. Bioinformatics. Vol 23:3113-3118 (2007)

[15] Pooja K. Stropea, and Etsuko N. Moriyama . Simple alignment-free methods for protein classification: A case study from G-protein-coupled receptors. Genomics Vol 89(5):602-612 (2007)

[16] Markus Wistrand1, Lukas Kll1 and Erik L.L. Sonnhammer . A general model of G protein-coupled receptor sequences and its application to detect remote homologs. Protein Science 15:509-521 (2006)

[17] S. Canu and Y. Grandvalet and V. Guigue and A. Rakotomamonjy. SVM and Kernel Methods Matlab Toolbox. Perception Systmes et Information, INSA de Rouen, Rouen, France (2005)