Functional Classification of G-Protein Coupled Receptors, Based on Their Specific Ligand Coupling Patterns

Burcu Bakir[1] and Osman Ugur Sezerman[2]

[1] School of Biology, Georgia Institute of Technology, Atlanta, USA
[2] Sabanci University, Istanbul, Turkey

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 remain as orphan sequences (the activating ligand is unknown). Therefore, classification methods for automated characterization of orphan GPCRs are imperative. In this study, for predicting Level 2 subfamilies of Amine GPCRs, a novel method for obtaining fixed-length feature vectors, based on the existence of activating ligand specific patterns, has been developed and utilized for a Support Vector Machine (SVM)-based 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 2 subfamilies of Amine GPCRs with a high predictive accuracy of 97.02% in a ten-fold cross validation test. The presented machine learning approach, bridges the gulf between the excess amount of GPCR sequence data and their poor functional characterization.

1 Introduction

G-Protein Coupled Receptors (GPCRs) are vital protein bundles with their key role in cellular signaling and regulation of various basic physiological processes. With their versatile functions in a wide range of physiological cellular conditions, they constitute one of the vastest families of eukaryotic transmembrane proteins [29]. In addition to the biological importance of their functional roles, their interaction with more than 50% of prescription drugs have lead GPCRs to be an excellent potential therapeutic target class for drug design and current pharmaceutical research. Over the last 20 years, several hundred new drugs have been registered which are directed towards modulating more than 20 different GPCRs, and approximately 40% of the top 200 synthetic drugs act on GPCRs [6]. Therefore, many pharmaceutical companies are involved in carrying out research aimed towards understanding the structure and function of these GPCR proteins. Even though thousands of GPCR sequences are known as a result of ongoing genomics projects [10], the crystal structure has been solved only for one GPCR sequence using electron diffraction at medium resolution (2.8 A) to date [15] and for many of the GPCRs the activating ligand is unknown, which are called orphan GPCRs [25]. Hence, based on sequence information, a functional classification method of those orphan GPCRs and new upcoming GPCR sequences is of great practical use in facilitating the identification and characterization of novel GPCRs.

Albeit laboratory experiments are the most reliable methods, they are not cost and labour effective. To automate the process, computational methods such as decision trees, discriminant analysis, neural networks and support vector machines (SVMs), have been extensively used in the fields of classification of biological data [21]. Among these methods, SVMs give best prediction performance, when applied to many real-life classification problems, including biological issues [30]. One of the most critical issues in classification is the minimization of the probability of error on test data using the trained classifier, which is also known as structural risk minimization. It has been demonstrated that SVMs are able to minimize the structural risk through finding a unique hyper-plane with maximum margin to separate data from two classes [27]. Therefore, compared with the other classification methods, SVM classifiers supply the best generalization ability on unseen data [30].

In the current literature, to classify GPCRs in different levels of families, there exist different attempts, such as using primary database search tools, e.g., BLAST [1], FASTA [20]. However, these methods require the query protein to be significantly similar to the database sequences in order to work properly. In addition to these database search tools, the same problem is addressed by using secondary database methods (profiles and patterns for classification), e.g., Attwood et al. have worked in particular on GPCRs in the PRINTS database [2] (whose data appeared in INTERPRO database [17]). Hidden Markov Models [24], bagging classification trees [32] and SVMs [13], [31] are other methods that have been used to classify GPCRs in different levels of families. Karchin et al. conducted the most comprehensive controlled experiments for sequence based prediction of GPCRs in [13] and showed that SVMs gave the highest accuracy in recognizing GPCR families. Whereas, 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. In [13], it is also pointed out that the SVM performance could be further increased by using feature vectors that encode only the most relevant features, 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 this paper, we address Level 2 subfamily classification of Amine GPCRs problem by applying Support Vector Machine (SVM) technique, using a novel fixed-length feature vector, based on the existence of activating ligand specific patterns. We obtain discriminative feature vectors by utilizing biological knowledge of the Level 2 subfamilies' transmembrane topology and identifying specific patterns for each Level 2 subfamily. Since these specific patterns carry ligand binding information, the features obtained from these patterns are more relevant features than amino acid and dipeptide composition of GPCR sequences, which in turn improves the accuracy of GPCR Level 2 subfamily classification. Applying our method on Amine Level 1 subfamily of GPCRs [10], we have shown that the classification accuracy is increased compared to the previous studies at the same level of classification.

2 Background

G-Protein Coupled Receptor Database (GPCRDB) information system organizes the GPCRs into a hierarchy of classes, Level 1 subfamilies (sub-families), Level 2 subfamilies (sub-sub-families), and types, based on the pharmacological classification of receptors [10]. A simplified view of GPCR family tree is presented in Figure 1. Since the binding of GPCRs into their specified ligands is important for drug design purposes, GPCRDB defines the classifications chemically (according to which ligands the receptor binds, based on the experimental data), rather than by sequence homology [13]. For class discrimination, generalization of the features shared by a diverse group of examples is required. Whereas, for subfamily discrimination, only the examples, that differ slightly, should be grouped together. Therefore, for GPCR subfamily classification problem, which is also related to GPCR function prediction, the ligand type that GPCR binds is more crucial than it is for GPCR class discrimination.

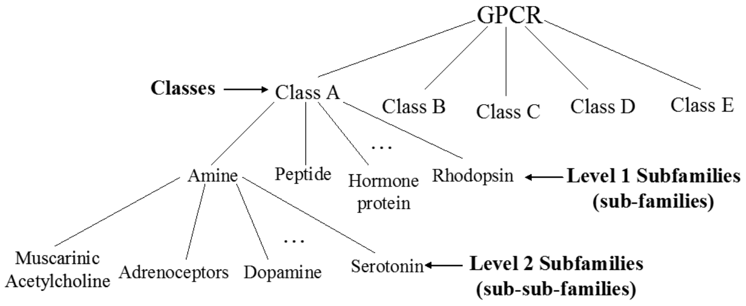

Fig. 1. Portion of GPCR family tree showing the five main classes of GPCRs and some subfamily members, based on the GPCRDB information system [10].

3 Materials and Methods

3.1 Benchmark Data

For GPCR function prediction from sequence information, subfamily recognition is more important than class recognition [13]. As mentioned before, subfamily recognition requires the knowledge of ligand coupling information of the receptor proteins. It is claimed that according to the binding of GPCRs with different ligand types, GPCRs are classified into at least six different families [9]. Among the sub-families in GPCRDB, Amine Level 1 subfamily of class A GPCRs is classified into seven sub-sub-families: (i) Muscarinic Acetylcholine, (ii) Adrenoceptors, (iii) Dopamine, (iv) Histamine, (v) Serotonin, (vi) Octopamine, (vii) Trace amine Level 2 subfamilies, according to the March 2005 release (9.0) of GPCRDB (Horn et al., 1998). Therefore, the correlation between sub-family classification and the specific binding of GPCRs to their ligands can be computationally explored for Level 2 subfamily classification of Amine Level 1 subfamily. Moreover, compared to the other classes, since Class A dominates by accounting for more than 80% of sequences as March 2005 release (9.0) of GPCRDB [10], it is the best studied class among different GPCRs. Thus, we will be able to compare our work with the previous studies. We use the same dataset, as that of Elrod and Chau, for Amine Level 1 subfamily GPCR sequences in GPCRDB, belonging to one of Acetylcholine, Adrenoceptor, Dopamine, Serotonin sub-sub-families, which have enough entries inside as a statistically significant training set, as shown in Table 1. The GPCR sequences in this dataset were extracted through the website http://www.expasy.org (SWISS-PROT database, Release 46.4, 2005) and fixed-length feature vectors are created for each sequence as it is explained in the next section.

Table 1. Summary of 168 Class A, Amine GPCRs, classified into four Level 2 Subfamilies as shown in [9].

Level 2 Subfamilies	Number of Sequences
Acetylcholine	31
Adrenoceptor	44
Dopamine	39
Serotonin	54
TOTAL	168

3.2 Fixed-Length Feature Vector Creation

Since protein sequences are of variable length, for classification, these sequences should be converted into fixed-length feature vectors. In order to obtain those fixed-length feature vectors, which also carry ligand specificity information, we followed a three step procedure as outlined in Figure 2. The first step, i.e., Topology Prediction step, aims to extract extracellular loop regions of the GPCR sequences since ligands couple to the outer loops of the GPCRs. So as to force fixed-length feature vectors to encode only biologically relevant features, activating ligand specificity information is taken into account. For this purpose,

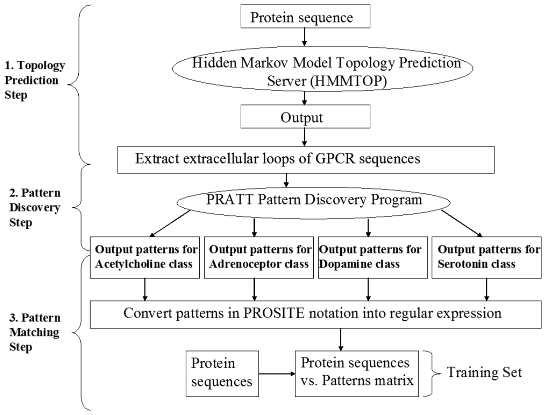
1. Topology Prediction Step
2. Pattern Discovery Step
3. Pattern Matching Step
Protein sequence
Hidden Markov Model Topology Prediction Server (HMMTOP)
Output
Extract extracellular loops of GPCR sequences
PRATT Pattern Discovery Program
Output patterns for Acetylcholine class
Output patterns for Adrenoceptor class
Output patterns for Dopamine class
Output patterns for Serotonin class
Convert patterns in PROSITE notation into regular expression
Protein sequences
Protein sequences vs. Patterns matrix
Training Set

Fig. 2. Flow chart for fixed-length feature vector creation.

conserved patterns, which are specific to each sub-sub-family of GPCR sequences are found in extracellular GPCR sequences in step two, i.e., Pattern Discovery step. In third step, i.e., Pattern Matching step, the existence of those activating ligand specific (specific for a sub-sub-family) patterns is checked. So that we integrate the coupling specificity of GPCRs into their ligands knowledge into our novel fixed-length feature vectors. Details of the three steps (Topology Prediction, Pattern Discovery, and Pattern Matching) for fixed-length feature vector creation are described below.

Topology Prediction Since transmembrane (TM) topology pattern is shown to be well conserved among GPCRs that have the same function [19], for the 168 GPCR sequences in Elrod and Chau's dataset, TM topology is checked. For topology prediction, Hidden Markov Model for Topology Prediction (HMMTOP) server, which accurately predicts the topology of helical TM proteins, is used [26]. In order to segment amino acid sequences into membrane, inside, outside parts, HMMTOP method utilizes HMMs in a way that the product of the relative frequencies of the amino acids of these segments along the amino acid sequence is maximized. This shows that the maximum of the likelihood function on the space of all possible topologies of a given amino acid sequence, correlates with the experimentally established topology [25].

Following topology prediction, extracellular loop sequences are extracted for each 168 GPCR sequences, based on the fact that ligands couple to extracellular loops of GPCRs and we are interested in the relation between ligand specificity of GPCRs and GPCR sub-sub-family classification.

Pattern Discovery In the second step of the fixed-length feature vector creation, for each sub-sub-family of GPCR sequences, flexible patterns that are conserved in the extracellular loop of that particular sub-sub-family GPCR sequences are found by using Pratt 2.1., flexible pattern discovery program [4]. Due to the flexibility of the Pratt patterns, they include ambiguous components, fixed and flexible wildcards, in addition to their identity components [12]. Hence, Pratt patterns are described using PROSITE notation [4].

Pratt finds patterns matching a subset of the input sequences. This subset is defined by "Min Percentage Seqs to Match (MPSM)" parameter, which defines the minimum percentage of the input sequences that should match a pattern. This threshold is set to 50% and 75% in this study in order not to allow for some very specific patterns that are not general to all GPCR sequences in any sub-sub-family. This can also be thought as a precaution to prevent overfitting problem. For each class of GPCRs, 50 conserved patterns are identified by two different MPSM parameters (50 and 75).

Pattern Matching The final step for creating fixed-length feature vectors is to check for the existence of every activating ligand specific pattern in each outer GPCR sequence. In order to check the existence of the flexible Pratt patterns, all patterns in PROSITE notation are converted into regular expression form and then they are searched within 168 extracellular GPCR sequences. Consequently, by taking activating ligand specific pattern existence information into account, each GPCR sequence is represented with a vector in the 200 dimensional space (50 patterns multiplied by 4 output classes).

$$\mathbf{G_k} = (G_{k,1}, G_{k,2}, \ldots, G_{k,200}) \tag{1}$$

where $G_{k,1}$, $G_{k,2}$ $G_{k,200}$ are the 200 components of activating ligand specific pattern inclusion for the k^{th} extracellular GPCR sequence G_k. Note that if the k^{th} extracellular GPCR sequence has the pattern j, then $G_{k,j}$=1 and if the k^{th} extracellular GPCR sequence does not have the pattern j, then $G_{k,j}$=0, where j=1, 2, ... 200.

Writing down each fixed-length feature vector, G_k, in a new row, we obtain a $G_{k,j}$ matrix, where k=1, 2, ... 168; j=1, 2, ... 200. After insertion of the sub-sub-family labels for each of the GPCR sequences into the zeroth dimension of each G_k vector ($G_{k,0}$), the matrix corresponds to a training set. So that k=0, 1, 2, ... 168, where $G_{k,0}$ is 1, 2, 3 or 4, since four sub-sub-families are defined for this classification problem. Note that these 4 class output labelling (1, 2, 3, 4) does not imply any relationship between classes.

We have also created a second fixed-length feature vector, by using the best 10 patterns among the 50 patterns based on significance scores assigned by the Pratt program from each sub-sub-family. Using a similar representation, G_k is denoted in 40 dimensional space (10 patterns multiplied by 4 output classes), where j=1, 2, ... 40. A $G_{k,j}$ matrix is formed (similar to above), where k=1, 2, ... 168 and j=1, 2, ... 40 corresponding another training set.

As a result, four training sets (two training sets with 50 MPSM parameter, for j up to 40 or 200; another two with 75 MPSM parameter, for j up to 40 or 200) are created to produce a classifier using Support Vector Machines, as mentioned in detail below.

3.3 Classification using Support Vector Machine (SVM)

The efficiency of SVMs for classification problems made them applicable in many real-world applications, including biological issues such as: protein classification and gene expression data analysis. SVM-based method for classification of subfamilies of GPCRs, is first developed by Karchin et al. [13]. When applied to the problem of discriminating both Level 1 and Level 2 subfamilies of GPCRs, SVMs are shown to make significantly fewer errors of both false positive and false negative than WU-BLAST and SAM-T2K profile HMMs [13]. For these reasons, we selected to use SVMs for GPCRs' Level 2 subfamily classification problem.

4 Experiments

Since we are interested in the classification of Amine Level 1 sub-family into four Level 2 subfamilies, we are facing with a multi-class classification problem. We use LIBSVM software [7], which deals with multi-class classification problem implementing "one-against-one" approach. As suggested in [11], to be able to get satisfactory results, some preprocesses are performed before building a classifier using LIBSVM. Preprocesses, that are performed in this study, can be summarized in two headlines: i) Choice of Kernel function, ii) Grid search combined with cross-validation for parameter tuning.

4.1 Choice of Kernel Function

Among linear, polynomial, radial basis function (RBF) and sigmoid Kernel functions, RBF kernel is a reasonable first choice as stated in [11], [14], [16]. Therefore, grid search and parameter tuning is done on RBF kernels. However, results obtained by using those four kernels are compared with parameter tuned RBF kernel at the end.

4.2 Cross-validation and grid search

In order to get better accuracy using RBF kernel for SVM classification, penalty parameter of error term, C, and γ parameter, which is specific to RBF kernel, should be tuned. Grid search procedure identifies the best (C, γ) pair, so that using these parameters the classifier (model) can accurately predict unseen test data [11]. Since the accuracy on test data also depends on the examples in the test data, cross validation is a better choice to tune (C, γ) parameters and select the best model that neither overfits nor underrepresents the training data. Compared to other advanced methods for parameter tuning, grid-search is straightforward, easy to implement and its computational time is not much more than advanced methods. Additionally, since each (C, γ) is independent, it can be easily parallelized. During grid search, it is recommended to try exponentially growing sequences of (C, γ) to identify good parameters [11].

To be able to solve our multi-class SVM classification problem, for each of our four training sets, grid search is performed for $C = 2^{-5}, 2^{-4}, \ldots, 2^{10}$ and $\gamma = 2^{5}, 2^{4}, \ldots, 2^{-10}$. Figure 3 shows the grid search with 10-fold cross validation for the training data with 200 attributes and 50 MPSM parameter. As it is seen in Figure 3, highest cross-validation accuracy is reached when γ=2^{-4} and C=(2^0, 2^{10}). After two preprocessing steps, as mentioned above, we build our classifiers using RBF kernel with best (C, γ) parameter pair, which is specific to the training set. Since we do not have a test set, and the number of examples in the training set is not big enough to separate into two, 10-fold cross-validation is done for each of the four training sets. Combining our biologically relevant fixed-length feature vector definition with a robust kernel, RBF, and parameter tuning with a grid search technique shows promising results, which is analyzed more in detail in the next section.

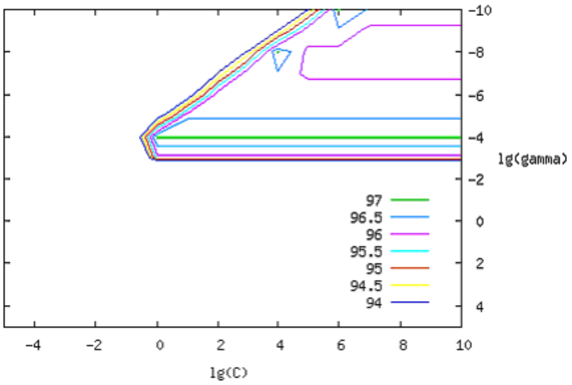

Fig. 3. Coarse grid search on C and with 10-fold cross validation for the training data with 200 attributes and 50 MPSM parameter. Highest cross-validation accuracy is obtained when γ=2^{-4} and C=(2^0, 2^{10}).

5 Results

As mentioned before, in addition to the SVM classification with parameter tuned RBF kernel, other three standard kernel functions are tested as well (with their default parameters) on our four training data using 10-fold cross validation. Results for each experiment are summarized in Table 2.

Classification with RBF kernel with parameter tuning clearly outperforms other kernel functions in all cases. Since linear kernel is the specialized form of RBF kernel, results obtained with these two kernels without parameter tuning are quite close. Although, the classification accuracy with 200 and 40 attributes are so close, accuracy with 200 attributes are consistently better than with 40 attributes. The probable reason behind this observation is that 40 attributes are not enough to represent the examples (more attributes are needed to discriminate between data points), or those chosen 40 attributes do not correctly reflect the data points. In contrast to the strict domination of 200 attributes over 40 attributes, there is no such a relationship between training data with 50 MPSM parameter and 75 MPSM parameter. While sometimes one performs better, it is vice versa (e.g. results of RBF Kernel and RBF* Kernel in Table 2). Lower accuracy for the training data with 75 MPSM parameter is caused by overfitting, which decreases accuracy at the end whereas with 50 MPSM parameter, patterns that are conserved in at least 50% of the data can not represent overall data.

In this paper, for the classification of Level 2 Subfamily of Amine GPCR's, 97.02% prediction accuracy is achieved in a ten-fold cross validation. Compared to the existing GPCR functional classification methods, for the classification of Level 2 Subfamily of Amine GPCR's, our result is superior to the SVM method with a simpler fixed-length feature vector definition and no parameter tuning, where prediction accuracy is 94.01% [31] and covariant discriminant algorithm, where prediction accuracy is 83.23% [9]. In another study, Ying et al. performs classification for both sub-family and sub-sub-family levels of GPCRs using bagging classification tree [32] . For sub-sub-family level, using the same dataset [9], our prediction accuracy in a ten-fold cross validation (97.02%) is higher than their prediction accuracy obtained in ten-fold cross validation (82.4%). More extensive comparison with previous studies is presented in the following section.

Table 2. Results for four different training sets, as explained in the text, using four different kernel functions and RBF kernel with parameter tuning (RBF*), with 10-fold cross-validation.

# of Attributes	MPSM Parameter	Linear Kernel	Polynomial Kernel	Sigmoid Kernel	RBF Kernel	RBF* Kernel
200	75	94.0476	48.8095	91.6667	91.6667	95.2381
200	50	96.4286	32.1429	86.3095	90.4762	97.0238
40	75	89.2857	47.0238	84.5238	86.3095	92.8571
40	50	92.8571	32.1479	84.5238	85.7143	93.4524

6 Discussion

The difference of this study from previous studies can be emphasized in two main points:

i) Fixed-length feature vector creation: We developed a novel method for obtaining fixed-length feature vectors of SVM. The naive idea that using direct protein sequence information as feature vector can not be used in SVM classification since the sequence length is not fixed. Many studies [9], [32], [31] attempted this problem by defining a fixed-length feature vector based on the protein's amino acid composition. Following the representation in [8], each protein is represented by a vector, X_k, in 20 dimensional space, where each dimension corresponds to how many times that particular amino acid, which represents that specific dimension, occurred in those particular protein.

$$\mathbf{X_k} = (X_{k,1}, X_{k,2}, \ldots, X_{k,20}) \tag{2}$$

where $X_{k,1}$, $X_{k,2}$ ... $X_{k,20}$ are 20 components of amino acid composition for the k^{th} protein X_k. In addition to the amino acid composition, in some of the studies, fixed-length vector is obtained by dipeptide composition [3], which takes local order of amino acids into account, in addition to the information about the fraction of amino acids. The dipeptide composition of each protein is shown using fractions of all possible dipeptides, where fraction of dipeptide i is the ratio of the number of dipeptide i in the protein divided by the total number of all possible dipeptides, namely 400. Alternatively, each protein sequence can also be transformed into a fixed-length feature vector, in the form of Fischer score vector [13].

Since in this study, the effect of activating ligand specificity in Level 2 Subfamily classification of Amine GPCRs is emphasized, a new feature vector is built, based on this observation. In this regard, we have used the existence information of activating ligand specific patterns, as fixed-length feature vectors, in order to come up with a biologically meaningful and distinctive measure. Therefore, the superiority of our feature vector stems from the biological importance of ligand coupling specificity for Level 2 Subfamily classification of Amine GPCRs. By combining those feature vectors with a robust kernel function, and parameter tuning strategy, we come up with an accurate classification method.

ii) Classification level: Apart from the definition of the feature vector for SVM, the exact classification level that we concentrate on, has been attempted in a few previous studies. Ying and Yanda and Ying et al. attempted the classification problem in the same Level 2 Subfamily of Amine GPCRs by using SVMs with amino acid composition as feature vector [31] and bagging classification tree [32], respectively. Our difference with their work is based on our novel feature vector definition as it is mentioned above, which in turn significantly affects the prediction accuracy (from 82.4% to 97.02% and 94.01% to 97.02% respectively). Apart from these particular papers, most of the previous studies concentrate on Superfamily level or Level 1 Subfamily. Although Karchin et al. have done experiments by using hierarchical multi-class SVMs, on Level 2 Subfamily [13] , they combine Class A Level 2 Subfamilies with Class C Level 2 Subfamilies.

Performance results in this study are promising and outperform other competitive techniques that classify GPCRs at the same level, with a very high cross validation accuracy of 97.02%. This result is mainly due to the definition of our feature vectors, since compared studies do not take into account such conserved pattern information for proper functioning of the GPCR. As the importance of specific ligand binding into GPCRs and the hidden information behind this binding is pointed out previously [13], we realized the use of ligand specific coupling patterns for creation of fixed-length feature vectors, which answers the need for biologically relevant features. Using these vectors for SVM classification and doing grid search for model selection, the accuracy have further improved even with very few sequences. With such accurate and automated GPCR functional classification methods, we are hoping to accelerate the pace of identifying proper GPCRs to facilitate drug discovery especially for schizophrenic and psychiatric diseases. Therefore, one of our future goals is to automate the presented procedure and come up with an integrated environment to perform GPCR classification conveniently with many flexible options to the biological users, who are not experts on the topic.

References

1. Altshul, S. et al.: Basic local alignment search tool. J. Mol. Biol. **215** (1990) 403–410
2. Attwood, T.K. et al.: PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Research **31** (2003) 400–402
3. Bhasin, M. and Raghava, G.: GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic Acids Research **32** (2004) 383–389
4. Brazma, A. et al.: Discovering patterns and subfamilies in biosequences. Proceedings of the Fourth International Conference on Intellignent Systems for Molecular Biology (ISMB-96), AAAI Press (1996) 34–43
   Pratt 2.1 software is available at www.ebi.ac.uk/pratt
5. Byvatov, E. and Schneider, G.: Support vector machine applications in bioinformatics. Appl. Bioinformatics **2** (2003) 67–77
6. Chalmers, D.T. and Behan, D.P.: The Use of Constitutively Active GPCRs in Drug Discovery and Functional Genomics. Nature Reviews, Drug Discovery **1** (2002) 599-608
7. Chang, C.C. and Lin, C.J.: LIBSVM : a library for support vector machines. (2001) LIBSVM software is available at http://www.csie.ntu.edu.tw/ cjlin/libsvm
8. Chou, K.C.: A Novel Approach to Predicting Protein Structural Classes in a (ZO-l)-D Amino Acid Composition Space. PROTEINS: Structure, Function, and Genetics **21** (1995) 319–344
9. Elrod, D.W. and Chou, K.C.: A study on the correlation of G-protein-coupled receptor types with amino acid composition. Protein Eng. **15** (2002) 713–715
10. Horn, F. et al.: GPCRDB: an information system for G protein coupled receptors. Nucleic Acids Res. **26** (1998) 275–279
    Available at: www.gpcr.org/7tm
11. Hsu, C.W. et al.: A Practical Guide to Support Vector Classification. Image, Speech and Intelligent Systems (ISIS) Seminars. (2004)
12. Jonassen, I. et al.: Finding flexible patterns in unaligned protein sequences. Protein Sci. **4** (1995) 1587–1595
13. Karchin, R. et al.: Classifying G-protein coupled receptors with support vector machines. Bioinformatics **18** (2001) 147–159

14. Keerthi, S.S. and Lin, C.J.: Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Computation **15** (2003) 1667-1689
15. Krzysztof, P. et al.: Crystal Structure of Rhodopsin: A G- Protein-Coupled Receptor. Science **4** (2000) 739–745
16. Lin, H.T. and Lin, C.J.: A study on sigmoid kernels for SVM and the train ing of nonPSDkernels by SMO type methods. Technical report, Department of Computer Science and Information Engineering, National Taiwan University. (2003)
Available at http://www.csie.ntu.edu.tw/ cjlin/papers/tanh.pdf
17. Mulder, N.J. et al.: The InterPro Database - 2003 brings increased coverage and new features. Nucleic Acids Research **31** (2003) 315–318
18. Neuwald, A. and Green, P.: Detecting Patterns in Protein Sequences. J. Mol. Biol. **239** (1994) 698-712
19. Otaki, J.M. and Firestein, S.: Length analyses of mammalian g-protein-coupled receptors. J. Theor. Biol. **211** (2001) 77–100
20. Pearson, W. and Lipman, D.: Improved tools for biological sequence analysis. Proceedings of National Academic Science **85** (1988) 2444-2448
21. Quinlan, J.R.: C4.5; Programs for Machine Learning, Morgan Kauffman Publishers, San Mateo, CA (1988)
22. Sadka, T. and Linial, M.: Families of membranous proteins can be characterized by the amino acid composition of their transmembrane domains. Bioinformatics **21** (2005) 378–386
23. Schoneberg, T. et al.: The structural basis of g-protein-coupled receptor function and dysfunction in human diseases. Rev Physiol Biochem Pharmacol. **144** (2002) 143–227
24. Sreekumar, K.R. et al.: Predicting GPCR-G-Protein coupling using hidden Markov models. Bioinformatics **20** (2004) 3490–3499
Swiss-Prot database (Release 46.4, 2005) is available at http:// www.expasy.org
25. Tusndy, G.E. and Simon, I.: Principles Governing Amino Acid Composition of Integral Membrane Proteins: Applications to topology prediction. J. Mol. Biol. **283** (1998) 489–506
26. Tusndy, G.E. and Simon, I.: The HMMTOP transmembrane topology prediction server. Bioinformatics **17** (2001) 849–850
Available at: http://www.enzim.hu/hmmtop
27. Vapnik, V. The Nature of Statistical Learning Theory, Springer-Verlag, New York. (1995)
28. Vert,J.P. Introduction to Support Vector Machines and applications to computational biology.(2001)
29. Vilo, J. et al.: Prediction of the Coupling Specificity of G Protein Coupled Receptors to their G Proteins. Bioinformatics **17** (2001) 174–181
30. Yang, Z.R.: Biological applications of support vector machines. Brief. Bioinform. **5** (2004) 328–338
31. Ying, H. and Yanda, L.: Classifying G-protein Coupled Receptors with Support Vector Machine. Advances in Neural Network (ISNN 2004), Springer LNCS. **3174** (2004) 448–452
32. Ying, H. et al.: Classifying G-protein Coupled receptors with bagging classification tree. Computational Biology and Chemistry **28** (2004) 275–280

This article was processed using the LaTeX macro package with LLNCS style