SVM-Based Classification of Class C GPCRs from Alignment-Free Physicochemical Transformations of Their Sequences

Caroline König[1], Raúl Cruz-Barbosa[2,3], René Alquézar[1], Alfredo Vellido[1]

[1] Univ. Politècnica de Catalunya. Barcelona Tech, 08034, Barcelona, Spain
{ckonig, alquezar, avellido}@lsi.upc.edu
[2] Univ. Tecnológica de la Mixteca, 69000, Huajuapan, Oaxaca, México
rcruz@mixteco.utm.mx
[3] Institut de Neurociències. Univ. Autònoma de Barcelona, 08193, Barcelona, Spain
raul.cruz@uab.es

Abstract. G protein-coupled receptors (GPCRs) have a key function in regulating the function of cells due to their ability to transmit extracelullar signals. Given that the 3D structure and the functionality of most GPCRs is unknown, there is a need to construct robust classification models based on the analysis of their amino acid sequences for protein homology detection. In this paper, we describe the supervised classification of the different subtypes of class C GPCRs using support vector machines (SVMs). These models are built on different transformations of the amino acid sequences based on their physicochemical properties. Previous research using semi-supervised methods on the same data has shown the usefulness of such transformations. The obtained classification models show a robust performance, as their Matthews correlation coefficient is close to 0.91 and their prediction accuracy is close to 0.93.

Key words: pharmaco-proteomics, G-Protein coupled receptors, homology, transformation, supervised learning, support vector machines

1 Introduction

G protein-coupled receptors (GPCRs) are cell membrane proteins with a key role in regulating the function of cells. This is the result of their ability to transmit extracellular signals, which makes them relevant for pharmacology. This has led, over the last decade, to active research in the field of proteomics.

The functionality of a protein depends widely on its 3-D structure, which determines its ability for certain ligand binding. Currently, the 3-D structure is only fully determined for approximately a 12% of the human GPCR superfamily [7]. As an alternative, when the information about the 3-D structure is not available, the investigation of the functionality of a protein can be achieved through the analysis of its amino acid sequence, which is known and available in several public curated databases.

Much research on sequence analysis has focused on the quantitative analysis of their aligned versions, although, recently, alternative approaches using machine learning techniques for the analysis of alignment-free sequences have been proposed. In this paper we focus on the alignment-free analysis of class C GPCRs, which have become an important research target for new therapies for pain, anxiety and neurodegenerative disorders.

The reported experiments concern a publicly available GPCR dataset that was analyzed, in a previous study, with semi-supervised techniques [3] as a strategy for GPCR deorphanization. Here, we extend this work through the use of a supervised multi-class classification approach. In this previous work, the analysis of the alignmnent-free sequences entailed a transformation of the symbolic sequences into real-valued feature vectors on the basis of the physicochemical properties of their constituent amino acids. In this study, the same transformations are used, including the Auto-Cross Covariance (ACC) transformation [15] and a more simple one: the amino acid composition (AA). To these, we add the Mean Transformation [10]. Some of these transformations have been used in previous research, such as in [10], where they were used to classifiy the five major GPCR classes using Partial Least Square Regression, and in [9], to classify a benchmark protein database using SVMs.

As previously mentioned, the current study uses primarily SVMs as the supervised classification model of choice for each of the transformed datasets. SVMs have been reported to be a top-performing method for protein classification [6,9] and are often attributed a high discriminating power due to their ability to use non-linear kernel functions to separate the input data in higher dimensional spaces. Nevertheless, some studies [2] report better results using more simple models such as Decision Trees (DTs) and Naive Bayes (NB). For this reason, these two techniques are compared with SVMs in the current study.

The obtained SVM classification models show a robust performance in the reported experiments. This is assessed using multi-class accuracy and Matthews Correlation Coefficient (MCC). The best results are obtained with the ACC-transformed dataset, achieving an MCC close to 0.91 and a prediction accuracy close to 0.93. GPCR subtype-specific results are also reported.

2 Materials

2.1 Class C GPCRs

GPCRs are cell membrane proteins with the key function of transmitting signals through it. Therefore, they are of special relevance in pharmacology. The GPCRDB [4], a popular database of GPCRs, divides the GPCR superfamily into five major classes (A to E) based on the ligand types, functions, and sequence similarities. The current study concerns class C of these receptors. This class has become an increasingly important target for new therapies, particularly in areas such as pain, anxiety, neurodegenerative disorders and as antispasmodics. They are also important from structural and mechanistic grounds. Whereas all

GPCRs are characterized by sharing a common seven transmembrane helices (7TM) domain, responsible for G protein activation, most class C GPCRs include, in addition, an extracellular large domain, the Venus Flytrap (VFT) and a cystein rich domain (CRD) connecting both [11].

Class C is further subdivided into seven types: Metabotropic glutamate (mGluR), Calcium sensing (CaSR), GABA-B, Vomeronasal (VN), Pheromone (Ph), Odorant (Od) and Taste (Ta). The investigated dataset consists of class C GPCR sequences obtained from GPCRDB[4], version 11.3.4 as of March 2011. A total of 1,510 sequences belonging to the seven types included: 351 *mGluR*, 48 *CaSR*, 208 *GABA-B*, 344 *VN*, 392 *Ph*, 102 *Od* and 65 *Ta*. The lengths of these sequences varied from 250 to 1,995 amino acids.

3 Methods

In this paper we use first of all SVMs for the classification of the alignment-free amino acid sequences and compare the results with those obtained by less complex techniques (DTs and NB). As the amino acid sequences have a variable length, one may apply sequence kernels to use them with SVMs or transform the sequence data to fixed-size vectors in order to use them with any supervised classifier, including non-kernel methods such as DTs and NB. The second approach has been followed in this work, which allows a comparison among different classifiers and also among different transformation methods.

3.1 Alignment-Free Data Transformations

- **Amino Acid Composition Transformation**: This transformation reflects the amino acid composition (AA) of the primary sequence, that is, the frequencies of 20 amino acids are calculated for each sequence (i.e., a $N \times 20$ matrix is obtained, where N is the number of items in the dataset).
- **Mean Composition Transformation**: This transformation applied in [10] first translates the amino acid sequence into physico-chemical descriptions, i.e. each amino acid is described by five z-scores [13]. In order to obtain a fixed-length representation of the sequence the average value of each z-score is calculated. This transformation generates a $N \times 5$ matrix.
- **Auto Cross Covariance Transformation**: The ACC transformation [8,15] is a more sophisticated transformation, which captures the correlation of the physico-chemical descriptors along the sequence. First the physico-chemical properties are represented by means of the five z-scores of each amino-acid as described by [13]. Then the Auto Covariance (AC) and Cross Covariance (CC) variables are computed on this first transformation. These variables measure respectively the correlation of the same descriptor (AC) or the correlation of two different descriptors (CC) between two residues separated by a lag along the sequence. From these, the ACC fixed length vectors can be

[4] http://www.gpcr.org/7tm/

obtained by concatenating the AC and CC terms for each lag up to a maximum lag, l. This transformation generates a $N\times\ (z^2 \cdot l)$ matrix, where $z = 5$ is the number of descriptors. In this work we use the ACC transformation for a maximal lag $l = 13$, which was found to provide the best accuracy for this dataset in [3].

3.2 Supervised Classification Techniques

Support Vector Machines (SVMs) [14] are complex classifiers with an ability to find a linear separation of instances in a higher dimensional space. DTs [12] predict class membership by examining the discriminative power of the attributes, whereas NB classifiers are probabilistic classifiers [5] that work under a simplifying assumption: attribute independence, that leads to efficient computation.

SVMs are based on the statistical learning theory first introduced in [14]. SVMs may map the feature vectors $x_i, i = 1, \ldots, N$, where $x_i \,\epsilon\, \mathbb{R}^n$ and N is the number of instances, into possibly higher dimensional spaces by means of a function ϕ . The objective is to find a linear separating hyperplane, which separates the feature vectors according to its class label with a maximal margin and minimizing the classification error ξ. The use of non-linear kernel functions allows SVMs to separate input data in higher dimensional spaces, which would not be separable with linear classifiers in the original input space.

The radial basis function (RBF) kernel, specified as $K(x_i, x_j) = e^{(-\gamma||x_i - x_j||)}$, is a popular non-linear kernel. Using it, the SVM needs to adjust two parameters through grid search: the error penalty parameter C and the parameter γ of the RBF function. Since our aim is to separate the seven subclasses of the class C GPCRs, this requires to extend the original two-class classification approach of SVMs to a multi-class classification approach. To that end, we have chosen the "one-against-one" approach to build the global classification model, which is implemented in the LIBSVM[5] library [1].

3.3 Criteria and Performance Measures

Two different measures were used to evaluate the test performance of the multi-class trained classifiers, namely the Accuracy (Accu), which is the proportion of correctly classified instances, and the MCC, which indicates how predictable the target variable is knowing the other variables: its value ranges from -1 to 1, where 1 corresponds to a perfect classification, 0 to a random classification and -1 to complete misclassification.

For the individual (binary) classification of each subtype, we report the MCC and two common measures: Precision and Recall. The former is the ratio of cases belonging to a class that are correctly classified to the cases predicted to belong to such class, whereas the later is the ratio of cases belonging to a class that are correctly classified to the cases that actually belong to that class.

[5] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

4 Experiments

4.1 SVM Model Selection

The SVM classification models are built upon the three transformed data sets. and involve the following processing steps:

1. Preprocessing of the dataset: Standardization of the data so that the mean is 0 and standard deviation is 1.
2. Splitting of the dataset into 5 stratified folds and applying 5-fold cross validation (5-CV) for the following steps:
   (a) Use the current training set for a parameter grid-search varying the parameters C and γ in a given range.
      i. For each combination of C and γ, determine the average classification accuracy using an inner 5-CV and update the parameters C and γ providing the best result.
      ii. Train an SVM model using the selected parameters C and γ and the current training set.
   (b) Classify the current test set with the SVM model obtained in the previous step recording the classification metrics aforementioned.
3. Calculate the mean value of the classification metrics recorded during step 2.b over the five outer iterations.

In our experiments we measure the Accuracy and MCC at the global level and the Precision, Recall and MCC at class level. The reported measures are the mean values of the respective metric over the five iterations of the (outer) 5-CV. At each iteration the aforementioned metrics are recorded for the SVM trained with the best parameters C and γ found in the corresponding grid search.

4.2 Model Selection Results

Table 1 shows details, for the three transformed datasets, of the grid searches conducted to find the optimal parameters C and γ of the RBF-SVM: the range of the tested parameters C and γ, the combination of parameters found to have the best performance, and the corresponding mean accuracy and MCC values on the test sets. The reported results of the grid search in Table 1 were confirmed with subsequent grid searches in smaller ranges of the parameters.

Table 1: Model selection results

DATA	**RANGE C**	**RANGE γ**	**PARAMETERS**	**Accu**	**MCC**
AA	1 to 16 (step +1)	2^{-5} - 2^{5} (step $\times 2$)	C=[2,8] , γ=2^{-4}	0.88	0.84
MEAN	1 to 16 (step +1)	2^{-5} - 2^{5} (step $\times 2$)	C=2 , γ=1	0.68	0.59
ACC	1 to 16 (step +1)	2^{-10} - 2^{5} (step $\times 2$)	C=[2,8] , γ=2^{-9}	0.93	0.91

4.3 Results and Discussion

The best classification results are found for the ACC transformed dataset using SVM classifiers (see Table 2 for a summary), achieving an accuracy of 0.93 and an MCC value of 0.91. The results obtained both for the ACC and the AA transformed datasets are consistent with those obtained with semi-supervised techniques in [3], where the ACC dataset also outperformed the AA dataset. Regarding classifier selection, SVM clearly outperforms DTs and NB for all three datasets (see Table 2 and Figure 1 for a comparison).

Table 2: Accuracy and MCC according to dataset and classifier

	SVM		**DT**		**NB**	
DATA	**Accu**	**MCC**	**Accu**	**MCC**	**Accu**	**MCC**
AA	0.88	0.84	0.74	0.67	0.72	0.65
MEAN	0.68	0.59	0.61	0.51	0.58	0.46
ACC	0.93	0.91	0.7	0.63	0.84	0.80

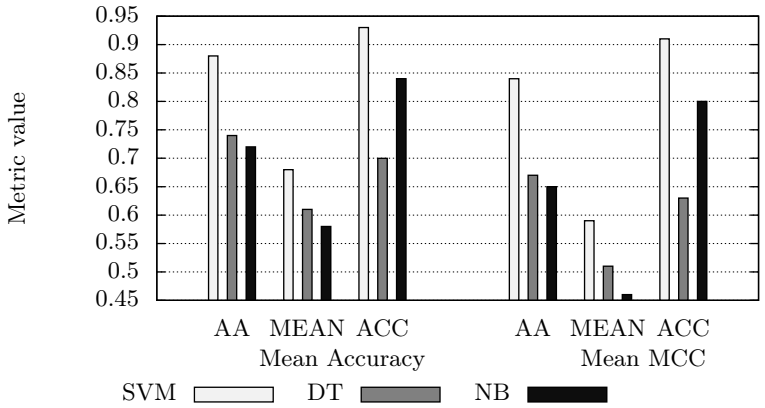

Fig. 1: Graphical representation of accuracy and MCC per dataset and classifier

Table 3 shows the classification results for the ACC-transformed dataset and the SVM classifier with greater detail at the per class level (these results correspond to a model with parameters C=2, γ=2^{-9}). The MCC value shows that classes *mGluR*, *CaSR*, *GABA-B* and *Ta* are very accurately discriminated from the other classes, having an MCC between 0.93 and 0.99. The prediction power of the classifier for classes VN, Ph and Od is clearly lower, with MCC values that range from 0.79 to 0.89.

As for the quality of the classifier, measured by the precision, it can be seen that it provides the most exact results for classes *CaSR*, *GABA-B* and *Ta*, as

Table 3: Per class results of SVM with the ACC data set

Class	**MCC**	**Precision**	**Recall**	**Type I error**	**Type II error**
mGluR	0.956	0.945	0.988	low	-
CaSR	0.933	1.000	0.877	-	high
GABA-B	0.986	0.990	0.985	-	-
VN	0.893	0.912	0.924	medium	medium
Ph	0.864	0.896	0.903	high	medium
Od	0.799	0.889	0.744	high	high
Ta	0.991	1.000	0.984	-	-

its precision gets very close to its maximum possible value. This metric shows that for classes *mGluR*, *Vn*, *Ph* and *Od* some type I classification errors (false positives) happen. Regarding the completeness of the classifier, measured by the recall, we see that it is most complete for classes *mGluR*, *GABA-B* and *Ta*, which means that nearly all real positives are correctly predicted. Classes *CaSR*, *Vn* and *Ph* have a lower recall, meaning that some type II errors (false negatives) happen for these classes. Class *Od* has a significantly lower recall than the other classes, what means that this class is most difficult to recognize.

Table 3 also shows an estimation of the quantity of type I and type II errors for each class. An analysis of these errors, by means of the confusion matrix, shows that the type II errors occur recurrently with a specific pattern for each class. For example, *Ph* are most frequently misclassified as *Vn* and less frequently as *mGluR* or *Od*. The existence of those patterns in the type II errors encourage an analysis of the class C dataset at the biochemical level in future work.

5 Conclusions

The supervised, alignment-free classification with SVMs of Class C GPCRs has been investigated in this paper. The experimental results have shown that the ACC transformed dataset has a clear advantage over the alternative transformations and that SVMs are best suited to the analysis of these data. The SVM classifiers built with this dataset and trained with the optimal parameters resulted highly accurate and discriminative. The per class results have shown some differences regarding the prediction power for some subclasses, which encourage the analysis of the less distinctive classes and the related classification errors in a future work at the biochemistry level.

Acknowledgments

This research is partially funded by Spanish research projects TIN2012-31377, SAF2010-19257, Fundació La Marató de TV3 (110230) and RecerCaixa 2010ACUP 00378. R. Cruz-Barbosa acknowledges Mexican council CONACYT for his post-doctoral fellowship.

References

1. C. Chang and C. Lin. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, 2011.
2. B. Cheng, J. Carbonell, and J. Klein-Seetharaman. Protein classification based on text document classification techniques. *Proteins: Structure, Function, and Bioinformatics*, 58(4):955–970, 2005.
3. R. Cruz-Barbosa, A. Vellido, and J. Giraldo. Advances in semi-supervised alignment-free classification of G protein-coupled receptors. In *Procs. of the International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO'13)*, pages 759–766, 2013.
4. F. Horn, E. Bettler, L. Oliveira, F. Campagne, F. Cohen, and G. Vriend. GPCRDB: An information system for G protein-coupled receptors. *Nucleic Acids Res*, 26:294–297, 1998.
5. G. John and P. Langley. Estimating Continuous Distributions in Bayesian Classifiers. In *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann, 1995.
6. R. Karchin, K. Karplus, and D. Haussler. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, 18(1):147–159, 2002.
7. V. Katritch, V. Cherezov, and R. C. Stevens. Structure-Function of the G Protein Coupled Receptor Superfamily. *Annual Review of Pharmacology and Toxicology*, 53(1):531–556, 2013. PMID: 23140243.
8. M. Lapinsh, A. Gutcaits, P. Prusis, C. Post, T. Lundstedt, and J.E.S. Wikberg. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Science*, 11(4):795–805, 2002.
9. B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan. Using Amino Acid Physicochemical Distance Transformation for Fast Protein Remote Homology Detection. *PLoS ONE*, 7(9), 2012.
10. S. O. Opiyo and E. N. Moriyama. Protein Family Classification with Partial Least Squares. *Journal of Proteome Research*, 6(2):846–853, 2007.
11. J. P. Pin, T. Galvez, and L. Prezeau. Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. *Pharmacology & Therapeutics*, 98(3):325 – 354, 2003.
12. J. R. Quinlan. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, 16(3):235–240, 1993.
13. M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, and S. Wold. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *Journal of Medicinal Chemistry*, 41(14):2481–2491, 1998.
14. V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.
15. S. Wold, J. Jonsson, M. Sjörström, M. Sandberg, and S. Rännar. DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Analytica Chimica Acta*, 277(2):239 – 253, 1993.