# Advances in Semi-Supervised Alignment-Free Classification of G Protein-Coupled Receptors

Raúl Cruz-Barbosa[1,2⋆], Alfredo Vellido[3], and Jesús Giraldo[1]

[1] Institut de Neurociències and Unitat de Bioestadística, Univ. Autònoma de Barcelona, 08193, Barcelona, Spain
{Raul.Cruz,Jesus.Giraldo}@uab.es
[2] Univ. Tecnológica de la Mixteca, 69000, Huajuapan, Oaxaca, México
rcruz@mixteco.utm.mx
[3] Univ. Politècnica de Catalunya. Barcelona Tech, 08034, Barcelona, Spain
avellido@lsi.upc.edu

**Abstract.** G Protein-coupled receptors (GPCRs) are integral cell membrane proteins of great relevance for pharmacology due to their role in transducing extracellular signals. The 3-D structure is unknown for most of them, and the investigation of their structure-function relationships usually relies on the construction of 3-D receptor models from amino acid sequence alignment onto those receptors of known structure. Sequence alignment risks the loss of relevant information. Different approaches have attempted the analysis of alignment-free sequences on the basis of amino acid physicochemical properties. In this paper, we use the Auto-Cross Covariance method and compare it to an amino acid composition representation. Novel semi-supervised manifold learning methods are then used to classify the several members of class C GPCRs on the basis of the transformed data. This approach is relevant because protein sequences are not always labeled and methods that provide robust classification for a limited amount of labels are required.

**Key words:** pharmaco-proteomics, G Protein-coupled receptors, semi-supervised learning, manifold learning, sequence alignment

## 1 Introduction

G Protein-coupled receptors (GPCRs) are integral cell membrane proteins of great relevance for pharmacology due to their role in transducing a wide range of extracellular signals. In doing so, they regulate the function of most cells in living organisms. The first GPCR crystal structure, that of rhodopsin, was only fully-determined in 2000 [1], and it is only over the last five years that the structures of some other 16 distinct receptors (approximately a 12% of human

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

GPCR super-family and, importantly, all belonging to GPCR class A) have been solved [2]. An alternative to work on GPCR structural models, when the 3-D crystal structures are not available, is the investigation of their functionality by analysis of their amino acid sequences, which are well documented and of which there are publicly available databases. Much of the existing research uses aligned versions of these sequences. Sequence alignment allows the use of more conventional quantitative analysis techniques, but at the price of risking the loss of the relevant information contained in the discarded sequence fragments.

Recently, different approaches have attempted the analysis of alignment-free sequences on the basis of their transformation according to the amino acid physicochemical properties (for a recent review see, for instance, [3]). In this paper, we use one of them to transform the data sequences: it takes the primary amino acid sequences and translates them into real-valued vectors based on those properties, followed by a transformation of the data into a uniform matrix by applying an Auto-Cross Covariance (ACC) transform [4]. A further and very simple amino acid sequence transformation is also used, consisting on the frequencies of 20 amino acids (thus not considering the order of the sequence).

Probabilistic modelling and, specifically, statistical machine learning (SML) models, even if in widespread use [5], have only recently begun to be applied in proteomics and, in particular, to the analysis of GPCRs.

In this paper, semi-supervised SML generative models of the manifold learning family are applied to the analysis of alignment-free sequences of class C GPCRs. The classification of GPCRs into families or classes and these into types and subtypes may contribute to the advancement of drug design and to a better understanding of the molecular processes involved in receptor signalling, both in normal and pathological conditions. A semi-supervised approach is thus relevant due to the fact that protein sequences are not always labeled (as assigned to a given subtype) and methods that can provide robust classification even with a limited amount of labels are required.

The experimental results indicate that semi-supervised methods working on the physicochemical properties of alignment-free class C GPCR sequences yield quite accurate classification even for a limited amount of available type labels. Amongst these methods, semi-supervised Generative Topographic Mapping (SS-GTM) consistently yields the best accuracy results. The use of the ACC data transformation is also shown to provide the most accurate classification.

## 2 Materials

### 2.1 Class C GPCRs

Membrane receptors are proteins to which signalling molecules may attach. They are the first step in the process of external signalling, allowing the initiation of intracellular signalling cascades after specific ligand binding. GPCRs are the most abundant family of membrane-bound receptors. They signal through their interaction and subsequent activation of G proteins [6]. It has been reported that

more than 50% of drugs target only four gene families, from which almost a 30% correspond to GPCRs. For this reason, they have become the subject of a vast research effort from the pharmaceutical industry.

The GPCRDB [8], 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. Here, we are interested in the class C of these receptors. This family 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 of G protein activation, most class C GPCRs include, in addition, an extracellular large domain, the Venus Flytrap (VFT) and a cysteine rich domain (CRD) connecting both [7].

Class C is, in turn subdivided into 7 types: Metabotropic glutamate, Calcium sensing, GABA-B, Vomeronasal, Pheromone, Odorant and Taste.

### 2.2 Analyzed data

The investigated dataset consisted of a total of 1,510 class C GPCR sequences, obtained from GPCRDB[4], version 11.3.4 as of March 2011. They belong to seven subfamilies, including: 351 metabotropic glutamate, 48 calcium sensing, 208 GABA-B, 344 vomeronasal, 392 pheromone, 102 odorant and 65 taste. The lengths of these sequences varied from 250 to 1,995 amino acids.

## 3 Methods

### 3.1 Alignment-Free Data Transformations

In this paper we consider two alignment-free data transformations. The first simply reflects the amino acid composition (AAcomp) 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). The second representation is provided by the more sophisticated ACC transformation [4, 9]. For this, each sequence is first translated into physico-chemical descriptions by representing each amino acid with the five $z$-scales derived in [10], then the Auto Covariance (AC) and Cross Covariance (CC) variables are computed on the transformed sequences. The AC measures the correlation of the same descriptor, $d$, between two residues separated by a lag, $l$, along the sequence. The CC variable measures the correlation of two different descriptors between two residues separated by a lag along the sequence. From these, the ACC fixed lenght vectors can be obtained. First, the AC and CC terms are concatenated for each lag ($C(l_i) = [AC(l_i)\ CC(l_i)]$) and then the ACC is obtained, for a maximum lag, $l_{max}$, by concatenating the $C(l_i)$ terms, that is, $ACC(l_{max}) = [C(l_1), \ldots, C(l_{max})]$. Details can be found in [4, 11].

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

### 3.2 Semi-supervised Generative Topographic Mapping

GTM [12] is a latent variable model in which a sample of $K$ regularly-spaced points $k = 1, \ldots, K$ residing in a low-dimensional space are mapped into the usually high-dimensional observed data space, each of them defining a prototype point. This prototype $\mathbf{y}_k$ is the image of the former according to the mapping function that takes the form $\mathbf{y}_k = \mathbf{W}\Phi(\mathbf{u}_k)$, where $\Phi$ is a set of $M$ nonlinear basis functions $\phi_m$, and $\mathbf{W}$ is a matrix of adaptive weights that defines the specific characteristics of the mapping. The prototype vector $\mathbf{y}_k$ can be seen as a representative of those data points $\mathbf{x}_n$ which are closer to it than to any other prototype and, thus, can also be seen as a cluster centroid. GTM performs a type of vector quantization that is similar to that of Self-Organizing Maps.

The set of prototypes $\mathbf{y}_k$ belongs to an intrinsically low-dimensional smooth manifold that wraps around the observed data $X = \{\mathbf{x}_n\}_{n=1}^{N}$. In this way, GTM becomes a manifold learning method. If we assume that the observed data lie close to the manifold, the conditional distribution of the observed data variables, given the latent variables, $p(\mathbf{x}|\mathbf{u})$ can be described as a noise model:

$$p(\mathbf{x}|\mathbf{u}, \mathbf{W}, \beta) = (\frac{\beta}{2\pi})^{D/2} \exp\{-\frac{\beta}{2} \sum_{d=1}^{D} (x^d - y^d(\mathbf{u}))^2\}, \quad (1)$$

with variance $\beta^{-1}$. From this, we can integrate the latent variables out, obtaining the likelihood of the model, and use maximum likelihood to estimate the adaptive parameters. Details of this procedure can be found in [12].

In many real settings, and proteomics is a perfect example of this, class labels may not be readily available for all cases. If ultimately interested in the classification of cases, we are faced with a semi-supervised learning problem [13] in which missing case labels must be inferred on the basis of the available ones.

Recently, GTM was redefined in a semi-supervised setting [14] as SS-Geo-GTM. For this, and understanding the model prototypes and manifold as the elements of a proximity graph, existing label propagation algorithms [15, 16] were adapted to a variant of GTM (namely Geo-GTM) in which Euclidean distances were replaced by approximations of geodesic distances along the GTM manifold.

A label vector $\mathbf{L}_k \in [0, 1]^c$ (where $c$ are the classes) is associated to each Geo-GTM prototype $\mathbf{y}_k$. The weights of the edges are derived from the graph distances $d_g$ between prototypes. The edge weight between nodes $k$ and $k'$ is calculated as $w_{kk'} = \exp(-d_g^2(k, k')/\sigma^2)$. The available label information of $\mathbf{x}_n \in X$ with class assignment $c(\mathbf{x}_n) = C_t \in \{C_1, \ldots, C_c\}$ is used to fix the label vectors of the prototypes to which they are assigned, so that $L_{k,j} = 1$ if $j = t$, and $L_{k,j} = 0$ otherwise. Unlabeled prototypes will then update their label by propagation, according to $\mathbf{L}_k^{new} = \sum_{k'} w_{kk'}\mathbf{L}_{k'} / \sum_{k'} w_{kk'}$. Unlabeled data items are finally labeled by assignment to the class of highest prevalence on the label vector of the prototype $\mathbf{y}_k$ that bears the highest responsibility for them, according to $c(\mathbf{x}_n) = \arg\max_{C_j \in \{C_1, \ldots, C_k\}} L_{k,j}$.

A detailed description of SS-Geo-GTM can be found in [14], whereas a practical application to a problem in the field of neuro-oncology is described in [17].

## 4 Experiments

The experiments reported in this section use the SS-Geo-GTM and SS-GTM summarily described in the previous section, but also alternatively, and for comparison, a semi-supervised SVM for manifold learning (SS-SVMan, [18]), which is a variant of the widely used SVM in which manifold learning is adapted to the semi-supervised setting in such a way that the objective function is modified to accommodate manifold consistency and the hinge loss of class prediction (an approximation to misclassification error). The result is an SVM-like process. There are three parameters involved in the choice of the SS-SVMan model: $C$, $\gamma$, and $\rho$. The last one is a coefficient that guarantees the invertibility of an expression leading to the obtention of the objective function. The other two parameters, typical of an SVM, are chosen for our experiments as indicated in [18].

SS-Geo-GTM, SS-GTM and SS-SVMan were all implemented in MATLAB®. For the experiments reported next, the matrix $\mathbf{W}$ and the inverse variance $\beta$ in SS-Geo-GTM and SS-GTM were initialized according to a standard procedure described in [12], which ensures the replicability of the results.

The goal of the experiments is twofold. Firstly, we aim to gauge the influence of the two alignment-free amino acid sequence representations (described in section 3.1) in the semi-supervised classification of class C GPCR subfamilies. Secondly, we aim to compare the performance of the three semi-supervised models in terms of classification accuracy.

### 4.1 Results and Discussion

Since unaligned amino acid sequences have varying lengths and our semi-supervised methods use vectors of shared dimensionality as input, data from the seven subfamilies of class C GPCRs were first transformed according to the two alignment-free representations described in section 3.1. In order to improve the accuracy results, a data normalization (or standardization) process can be applied in such way that the columns of the data matrix have zero mean and unit standard deviation. Once the AAcomp and ACC transformations were applied to the data under analysis, two datasets were obtained by data normalization.

The figure of merit for the semi-supervised models is the average classification accuracy over 100 runs. Labels were available for all sequences in the sample extracted from the database. To evaluate the models in a semi-supervised setting, labels were removed (becoming *missing*) randomly. The class label availability was made to vary from a very extreme (1%) to a relaxed (30%) setting.

The average classification results for the dataset obtained using the AAcomp transformation are shown in Table 1. From these results, the simple SS-GTM is shown to outperform the other methods in the most extreme settings up to 10% labeled data availability, which means that the unsupervised nature of GTM-based models can help to discover the class structure in a better way when very few labeled data are available. On the contrary, when the label availability condition is relaxed, the SS-SVMan model outperforms the GTM-based methods, which means that the supervised nature of SVM-based models is likely to better

**Table 1.** Classification accuracy as an average percentage over one hundred runs (with its corresponding standard deviation) using the AAcomp representation. Superscript $DN$ indicates that data normalization pre-processing was applied.

| **Method** | **Percent of available labels** | | | | |
|---|---|---|---|---|---|
| | 1 | 5 | 10 | 20 | 30 |
| *SS-GTM*$^{DN}$ | 49.37 ± 5.26 | 68.42 ± 3.09 | 75.28 ± 1.84 | 80.55 ± 1.29 | 82.21 ± 1.05 |
| *SS-Geo* | 40.81 ± 4.23 | 61.06 ± 3.00 | 69.07 ± 2.25 | 76.52 ± 1.44 | 80.08 ± 0.99 |
| *SS-SVMan*$^{DN}$ | 43.78 ± 5.57 | 65.51 ± 3.22 | 74.38 ± 1.96 | 81.69 ± 1.17 | 85.84 ± 0.91 |

reveal the class structure only when enough labeled data (as much as 30% in this dataset) are available. As for the data normalization, very poor accuracy results (below 50%) were obtained for SS-SVMan with 30% of labeled data when non-normalized data were used. The same setting for SS-GTM slightly reduced its results to those of SS-Geo-GTM (all these results are not shown for the sake of brevity).

Next, and in order to apply the ACC transformation to GPCR sequences, each of them was first translated into physicochemical descriptions by representing each amino acid with the five $z$-scales derived by [10]. Then, the ACC fixed length vectors were computed. The ACC transformation uses two parameters: a maximum lag $L$ and a degree of normalization $p$, to be tuned. For this, the optimal parameters were experimentally chosen by investigating the impact on classification accuracy of multiple combinations of their values. Previous experiments, as in [4] and [11], have shown that the maximum lag is to be found in the range [1,160]. Following [11], we searched for $L$ in the range of 1 to 30. The $p$ parameter was set at different values, including: 0, 0.5 and 1.0. The average classification results were computed for SS-GTM using 30% of labeled data for each combination. As an illustration, the classification accuracy results for $p$ values of 0.5 and 1.0 are shown in Fig. 1 (the results for $p = 0$ are not stable, which suggests that a large $L$ is needed). It can be observed that $p = 0.5$ provides the best results. Classification accuracy of around 85% was achieved with a lag of 7 and results stabilized from a value of 13 onwards. For computational time expediency, a maximal lag of 13 was thus selected.

The average classification results for the dataset using the ACC representation with $L = 13$ and $p = 0.5$ are shown in Table 2. The performance of the analyzed methods follows the same tendency as in Table 1, but more pronounced in favour of GTM-based models this time. The results for SS-GTM and SS-Geo-GTM are very similar and these, in turn, clearly outperform the SS-SVMan using from 1% to 20% labeled data. SS-SVMan is competitive only when enough labeled data (30% in this case) are available.

The results in Tables 1 and 2 reflect the advantage of using the ACC representation, which yields, consistently, the best classification results. This is specially clear for the extreme semi-supervised settings (using 1, 5 and 10 percent of labeled data). On the other hand, the simple AAcomp representation becomes competitive when the label availability condition is relaxed to 20% and 30%,

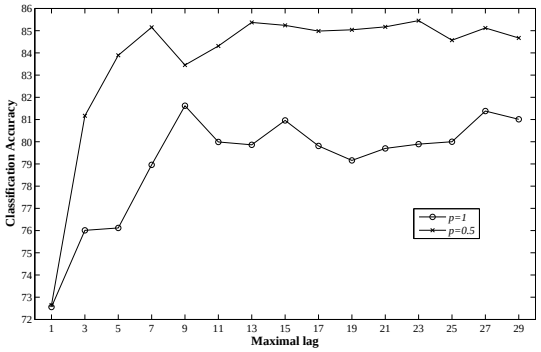


**Fig. 1.** Average classification accuracy over 100 runs for different maximal lags.

**Table 2.** Classification accuracy as an average percentage over one hundred runs (with its corresponding standard deviation) using the ACC representation. Superscript $DN$ indicates that data normalization preprocessing was applied.

| **Method** | **Percent of available labels** | | | | |
|---|---|---|---|---|---|
| | 1 | 5 | 10 | 20 | 30 |
| *SS-GTM* | 59.12 ± 6.56 | 76.42 ± 2.83 | 81.24 ± 2.25 | 84.29 ± 1.45 | 85.37 ± 1.26 |
| *SS-Geo* | 57.71 ± 5.40 | 75.21 ± 3.29 | 81.33 ± 2.02 | 84.09 ± 1.33 | 84.82 ± 1.34 |
| *SS-SVMan*$^{DN}$ | 31.84 ± 4.40 | 58.27 ± 3.18 | 71.02 ± 2.32 | 82.04 ± 1.41 | 87.29 ± 1.16 |

which means that it could be recommended for a first and fast semi-supervised analysis only when enough data are available. In summary, according to the obtained classification results, the ACC representation can be recommended for extreme semi-supervised settings; however, the AAcomp transformation could be applied for a first and fast analysis if enough labeled data were available.

## 5 Conclusions

The semi-supervised, alignment-free classification of Class C GPCRs has been investigated in this paper. The preliminary experimental results indicate that semi-supervised GTM-based models work well in situations of extreme class label scarcity, whereas semi-supervised SVM is competitive only when enough labeled data are available. The alignment-free GPCRs representations have shown a key role in the classification accuracy results. The physicochemical properties-based ACC representation has shown clear advantage in settings of extreme label scarcity, capturing discriminative characteristics from the sequences even in the near-absence of class information. The performance of the order-independent AAComp method also becomes competitive when enough class information is available, making it a viable first and fast alternative in these cases. Further research on other kinds of GPCRs sequence transformations is encouraged.

## References


1. Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., et al. Crystal Structure of Rhodopsin: a G Protein-Coupled Receptor. Science 289:73945 (2000)
2. Katritch, V., Cherezov, V., Stevens, R.C. Structure-Function of the G Protein-Coupled Receptor Superfamily. Annu Rev Pharmacol Toxicol. doi:10.1146/annurev-pharmtox-032112-135923 (2012)
3. Liu, B., Wang, X., Chen, Q., Dong, Q., Lan, X.: Using Amino Acid Physicochemical Distance Transformation for Fast Protein Remote Homology Detection. PLoS ONE 7(9):e46633 (2012)
4. Lapinsh, M., Gutcaits, A., Prusis, P., Post, C., Lundstedt, T., Wikberg, J.E.S.: Classification of G-protein Coupled Receptors by Alignment-Independent Extraction of Principal Chemical Properties of Primary Amino Acid Sequences. Protein Sci. 11:795805 (2002)
5. Hastie, T., Tibshirani, T., Friedman, J.: The Elements of Statistical Learning ($2^{nd}$ ed.). Springer, Heidelberg (2009)
6. Gilman A.G.: G Proteins: Transducers of Receptor-Generated Signals. Annu Rev Biochem. 56, 615–649 (1987)
7. Pin, J.P., Galvez, T., Prézeau, L.: Evolution, Structure, and Activation Mechanism of Family 3/C G-protein-coupled receptors. Pharmacol Ther. 98(3):325-54 (2003).
8. Horn F., Weare J., Beukers M.W., Horsch S., Bairoch A., Chen W., Edvardsen O., Campagne F., Vriend G.: GPCRDB: An Information System for G Protein-Coupled Receptors. Nucleic Acids Res. 26, 275–279 (1998)
9. Wold, S., Jonsson, J., Sjöström, M., Sandberg, M., Rännar, S.: DNA and Peptide Sequences and Chemical Processes Multivariately Modelled by Principal Component Analysis and Partial Least-Squares Projections to Latent Structures. Anal. Chim. Acta 277, 239–253 (1993)
10. Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., Wold, S..: New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. J. Med. Chem. 41, 2481–2491 (1998)
11. Otaki, J.M., Mori, A., Itoh, Y., Nakayama, T., Yamamoto, H.: Alignment-Free Classification of G-Protein Coupled Receptors using Self-Organizing Maps. J. Chem. Inf. Model 46, 1479–1490 (2006)
12. Bishop, C.M., Svensén, M., Williams, C.K.I.: GTM: The Generative Topographic Mapping. Neural Comput. 10(1), 215–234 (1998)
13. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. The MIT Press, MA(USA) (2006)
14. Cruz-Barbosa, R., Vellido, A.: Semi-Supervised Geodesic Generative Topographic Mapping. Pattern Recogn Lett. 31(3), 202–209 (2010)
15. Zhu, X., Ghahramani, Z.: Learning from Labeled and Unlabeled Data with Label Propagation. Technical report, CMU-CALD-02-107, Carnegie Mellon Univ. (PA) U.S.A. (2002)
16. Herrmann, L., Ultsch, A.: Label Propagation for Semi-Supervised Learning in Self-Organizing Maps. In: 6th International Workshop on Self-Organizing Maps (WSOM), Neuroinformatics Group, Bielefeld University, Germany (2007).
17. Cruz-Barbosa, R., Vellido, A.: Semi-Supervised Analysis of Human Brain Tumours from Partially Labeled MRS Information, using Manifold Learning Models. Int J Neural Syst. 21(1):17-29 (2011)
18. Wu, Z., Li, C.H., Zhu, J., Huang, J.: A Semi-Supervised SVM for Manifold Learning. In Procs. of the 18th International Conference on Pattern Recognition. IEEE Computer Society (2006)