Manifold learning visualization of Metabotropic Glutamate Receptors

Martha-Ivón Cárdenas [a,b,1], Alfredo Vellido [a,c] and Jesús Giraldo [b]

[a] *Llenguatges i Sistemes Informàtics, UPCatalunya 08034, Barcelona, Spain*
[b] *Institut de Neurociències and Unitat de Bioestadística, UAB, 08193, Bellaterra, Spain*
[c] *CIBER-BBN, Cerdanyola del Vallès, Spain*

Abstract. G-Protein-Coupled Receptors (GPCRs) are cell membrane proteins with a key role in biological processes. GPCRs of class C, in particular, are of great interest in pharmacology. The lack of knowledge about their 3-D structures means they must be investigated through their primary amino acid sequences. Sequence visualization can help to explore the existing receptor sub-groupings at different partition levels. In this paper, we focus on Metabotropic Glutamate Receptors (mGluR), a subtype of class C GPCRs. Different versions of a probabilistic manifold learning model are employed to comparatively sub-group and visualize them through different transformations of their sequences.

Keywords. G-Protein-Coupled Receptors, Metabotropic Glutamate Receptors, data visualization, Generative Topographic Mapping.

Introduction

The G-protein-coupled receptors (GPCRs) in the human genome form five main families (A to E) according to their similarity [9]. Class C GPCRs include metabotropic glutamate receptors (mGluRs), in which we focus our research. They are promising targets for the development of new therapeutic drugs.

The functionality of GPCRs is often studied from the 3-D structure of their sequences. As no complete crystal structure data is currently available for class C GPCRs, the investigation of their primary structure as amino acid (AA) sequences is necessary. The unaligned symbolic sequences are unsuitable for direct analysis, but many different sequence transformation techniques are available to overcome this limitation. In this study, we used two relatively simple ones: the first is AA composition (AAC [2]), which accounts only for the relative frequencies of appearance of the 20 AAs in the sequence. Recent analysis using semi-supervised and supervised classification [3,4] with this type of transformation showed that accuracy reaches an upper bound.The second choice is the *digram* transformation, which considers the frequencies of occurrence of any given pair

[1]*This research was partially funded by MINECO TIN2012-31377 and SAF2010-19257, as well as Fundació La Marató de TV3 110230 projects.

of AAs. They were used for the more general classification of class C GPCR sequences in [5], obtaining accuracies in the area of 93-94%.

The target of this study was exploratory mGluR sequence clustering and visualization as a preliminary but complementary step towards full-blown mGluR subtype classification (into their eight known subtypes). This was implemented using different variants of a nonlinear dimensionality reduction (NLDR) method: Generative Topographic Mapping (GTM [6]). This machine learning technique has previously been applied with success to the more general problem of class C GPCR visualization [7,8]. MGluR subtype visual discrimination was quantitatively assessed here using an entropy measure.

1. Materials and methods

1.1. Class C GPCR mGluR data

The GPCRDB [9] database of GPCRs divides them into five major classes (namely, A to E). The investigated class C data (from version 11.3.4 as of March 2011) include 351 mGluR sequences, in turn sub-divided into 8 subtypes (mGluR1 to mGluR8) plus a group of mGluR-like sequences. They are distributed as 33 cases of mGluR1, 26 mGluR2, 44 mGluR3, 23 mGluR4, 32 mGluR5, 15 mGluR6, 4 mGluR7, 98 mGluR8 and 76 mGluR-like. This 8 subtypes can also be grouped into 3 categories according to sequence homology, pharmacology and transduction mechanism: *group I* mGluRs include mGluR1 and mGluR5; *group II* includes mGluR2 and mGluR3; whereas *group III* includes mGluR4, 6, 7 and 8.

1.2. The basic GTM and Kernel GTM

The GTM [6] is a non-linear latent variable model of the manifold learning family that performs simultaneous data clustering and visualization through a topology-preserving generative mapping from the latent space in $\mathbb{R}^{\mathbb{L}}$ (with $L=2$ for visualization) onto the $\mathbb{R}^{\mathbb{D}}$ space of the observed data in the form $\mathbf{y}=\Phi(\mathbf{u})W$, where $\mathbf{y}$ is a D-dimensional vector, Φ is a set of M basis functions, $\mathbf{u}$ is a point in the visualization space and W is a matrix of adaptive weights w_{md}. The likelihood of the full model can be approximated and maximum likelihood methods can be used to estimate the adaptive parameters. Details can be found in [6] and elsewhere. The probability of each of the K latent points $\mathbf{u}_k$ for the generation of each data point $\mathbf{x}_n$, $p(k|\mathbf{x}_n)$, also known as a *responsibility* r_{kn}, can be calculated as part of the parameter estimation process. For data visualization, it is used to obtain a *posterior mode projection*, defined as x_n: $k_n^{mode}=\arg\max_{\{k_n\}} r_{kn}$, as well as a *posterior mean projection* $k_n^{mean}=\sum_{k=1}^{K} r_{kn}\mathbf{u}_k$. The standard GTM is used here to model and visualize the AAC- and digram-transformed unaligned sequences.

The kernel-GTM (KGTM) [10] is a kernelized version of the standard GTM that is specifically well-suited to the analysis of symbolic sequences such as those characterizing proteins. This is achieved by describing sequence similarity through a kernel function based on the mutations and gaps between sequences: $K(\mathbf{x},\mathbf{x}')=\rho\exp\left\{\nu\frac{\pi(\mathbf{x},\mathbf{x}')}{\pi(\mathbf{x},\mathbf{x})+\pi(\mathbf{x}',\mathbf{x}')}\right\}$ for sequences $\mathbf{x}$ and $\mathbf{x}'$; ρ and ν are prefixed parameters, and $\pi(\cdot)$ is a score function of common use in bioinformatics. Further details on these parameters can be found in [10]. KGTM is used here to model and visualize the multiple sequence alignment (MSA)-transformed sequences, using the *posterior mode projection*.

2. Results

The standard GTM visualization of the AAC- and digram-transformed mGluR sequences according to their *posterior mean projection* is shown in Fig.1.

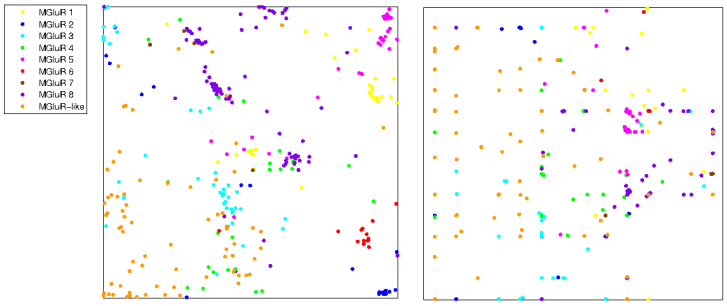

Figure 1. Visualization map of the standard GTM-based *posterior mean projection* of the mGluR AAC- (left) and digram-transformed (right) sequences. Different mGluR subtypes are identified by color, as in Fig. 2.

Given that, for KGTM, all the conditional probabilities (responsibilities) r_{kn} are sharply peaked around the latent points $\mathbf{u}_k$, the visualization of the mGluR is better and more intuitively represented by their *posterior mode projections* as shown in Fig.2.

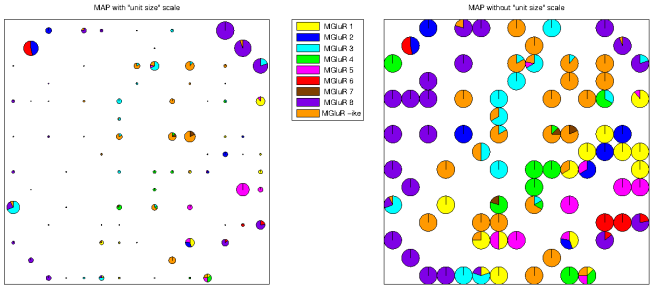

Figure 2. KGTM-based visualization of the mGluR subtypes through their *posterior mode projection*. Left) Individual pie charts represent sequences assigned to a given latent point and their size is proportional to the ratio of sequences assigned to them by the model. Each portion of a chart corresponds to the percentage of sequences belonging to each mGluR subtype. Right) The same map without sequence ratio size scaling, for better visualization.

An entropy-based measure, suitable for discrete clustering visualizations, was used to quantify the level of mGluR subtype overlapping: If map areas are completely subtype specific, entropy will be zero, whereas high entropies will characterize highly overlapping subtypes. For a given latent point k, the entropy is $S_k = -\sum_{j=1}^{C} p_{kj} ln p_{kj}$, where j is one of the $C = 9$ mGluR plus mGluR-like subtypes and $p_{kj} = \frac{m_{kj}}{m_k}$, where, in turn, m_k is the number of sequences in cluster k and m_{kj} is the number of subtype j sequences in cluster k. The total entropy of a given GTM map can thus be calculated as $S = \sum_{k=1}^{K} \frac{m_k}{N} S_k$, where $\frac{m_k}{N}$ is the proportion of mGluR sequences assigned to latent point k. The entropy results for the standard GTM representation of the transformed sequences are summarized in Table 1.

Table 1. Entropies for each of the 8 mGluR and mGluR-like subtypes, together with total entropy.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	Like	**Total**
AAC	0.35	0.41	0.69	0.65	0.55	0.19	0.41	0.33	0.48	0.31
digram	0.77	0.55	0.53	0.59	0.50	0.80	0.69	0.48	0.35	0.37
KGTM	0.50	0.65	0.51	0.47	0.53	0.57	0.64	0.27	0.50	0.33

3. Discussion

All GTM visualizations provide insights about the inner grouping structure of mGluRs. The first overall finding is that most subtypes show a reasonable level of separation, but none of them avoids subtype overlapping. Most subtypes also show clear inner structure. The differences between the AAC sequence mapping and its *digram* counterpart in Fig.1 are noticeable, although there are also clear coincidences, such as the neat separation of the heterogeneous mGluR-like sequences in the bottom-left quadrants of both maps, with mGlu3 located nearby. These differences indicate that the visual data representation is at least partially dependent on the type of sequence transformation. This is further corroborated by the KGTM visualization in Fig.2. The mapping differs in many ways from the previous ones, although many characteristics remain consistent. As stated in section 1.1, the 8 main mGluR subtypes are commonly grouped into 3 categories. The visualizations in Figs.1 and 2 provide only partial support to these categories.

The entropy measure described in the previous section provides us with a quantitative measure of subtype location specificity. The results in Table 1 are quite telling. First, because the overall entropy is not too dissimilar between transformations; despite this, the transformation yielding lowest entropy (highest level of subtype discrimination) is, unexpectedly, the simplest one: AAC, which does not even consider ordering in the AA sequence. It is clear, in any case, that subtype overlapping is substantial. Second, because the dependency of results on the type of sequence transformation is clearly confirmed.

References

[1] M.C. Lagerström and H.B. Schiöth, Structural diversity of G protein-coupled receptors and significance for drug discovery, *Nature Reviews Drug Discovery* **7** (2008), 339–357.

[2] M. Sandberg *et al.*, New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids, *Journal of Medicinal Chemistry* **41** (1998), 2481–2491

[3] R. Cruz-Barbosa, A. Vellido, and J. Giraldo, Advances in semi-supervised alignment-free classification of G-protein-coupled receptors. IWBBIO'13, Granada, Spain, pp. 759–766 (2013)

[4] C. König *et al.*, SVM-based lassification of class C GPCRs from alignment-free physicochemical transformations of their sequences. ICIAP 2013, LNCS 8158, pp. 336–343, (2013)

[5] C. König *et al.*, Finding class C GPCR subtype-discriminating n-grams through feature selection. PACBB 2014.

[6] C.M. Bishop, M. Svensén, and C.K.I. Williams, GTM: The Generative Topographic Mapping. *Neural Computation* **10** (1998), 215–234.

[7] M.I. Cárdenas, A. Vellido, I. Olier, X. Rovira, and J. Giraldo, Complementing kernel-based visualization of protein sequences with their phylogenetic tree, CIBB 2011, LNCS/LNBI 7548, 136–149 (2012)

[8] M.I. Cárdenas *et al.*, Exploratory visualization of misclassified GPCRs from their transformed unaligned sequences using manifold learning techniques. IWBBIO 2014, 623–630 (2014)

[9] B. Vroling, *et al.*, GPCRDB: information system for G protein-coupled receptors. *Nucleic Acids Research* **39**, suppl 1 (2011) D309–D319

[10] I. Olier *et al.*, Kernel Generative Topographic Mapping. ESANN 2010, 481–486 (2010)