Exploratory Visualization of Misclassified GPCRs from Their Transformed Unaligned Sequences Using Manifold Learning Techniques

Martha I. Cárdenas[1,2]*, Alfredo Vellido[1], Caroline König[1], René Alquézar[1], and Jesús Giraldo[2]

[1] Departament de Llenguatges i Sistemes Informàtics,
Universitat Politècnica de Catalunya 08034, Barcelona - Spain
[2] Institut de Neurociències, Unitat de Bioestadística,
Universitat Autònoma de Barcelona, 08193, Bellaterra (Barcelona) - Spain

Abstract. Class C G-protein-coupled receptors (GPCRs) are cell membrane proteins of great relevance to biology and pharmacology. Previous research has revealed an upper boundary on the accuracy that can be achieved in their classification into subtypes from the unaligned transformation of their sequences. To investigate this, we focus on sequences that have been misclassified using supervised methods. These are visualized, using a nonlinear dimensionality reduction technique and phylogenetic trees, and then characterized against the rest of the data and, particularly, against the rest of cases of their own subtype. This should help to discriminate between different types of misclassification and to build hypotheses about database quality problems and the extent to which GPCR sequence transformations limit subtype discriminability. The reported experiments provide a proof of concept for the proposed method.

Keywords: G-Protein Coupled Receptors; Data Visualization; Manifold Learning; Unaligned Sequence Analysis; Phylogenetic Trees

1 Introduction

G-protein-coupled receptors (GPCRs) are cell membrane proteins of great relevance to biology due to their role in transducing extracellular signals. Importantly, more than a third of all drugs approved by the US Food and Drug Administration over the last three decades actually target GPCRs [1], which makes them the object of large-scale research in the pharmaceutical industry.

The current study focuses on class C GPCRs, which have recently become an increasingly important target for new therapies [2]. The 3-D structure of proteins is usually the key to the understanding of their function. Despite intensive research efforts, no complete class C 3-D structure has yet been unraveled, which means that we are limited to the investigation of their primary structure:

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

the amino acid (AA) sequences, of which several databases are publicly available. The sequence diversity of class C GPCRs makes them a challenging target for classification, which can be performed at many different levels of detail (for GPCRs as a whole, hierarchical classification was recently achieved at seven subtyping levels, starting from GPCR vs. non-GPCR [3]). The correct classification of class C GPCR sequences into their subtypes is the basis of the current study.

Many existing GPCR classification systems use aligned versions of sequences, risking the loss of relevant information contained in discarded sequence fragments. There are different ways to bypass this limitation and use full alignment-free sequences for classification. Given the exploratory goal of this study, we focus on a very simple AA sequence transformation that considers only the relative frequencies of appearance of the 20 AAs in the sequence (thus ignoring the sequential order). Recent analysis using semi-supervised classification of class C GPCRs [5] with this type of transformation showed that accuracy reaches an upper bound (between 80-85%) that it is not significantly increased when more sophisticated physico-chemical transformations of the sequences are applied (never reaching 90%). Although the simplicity of this transformation also risks losing relevant information, recent experiments using supervised Support Vector Machine (SVM) classifiers [6] yielded best results in the area of 88%. A detailed review about this type of classification can be found in [7].

To investigate this classification bound, we propose in this study a method that combines GPCR classification with multivariate data (MVD) visualization using the unaligned transformed sequences as a starting point. Visualization is used here as an exploratory Data Mining tool, facilitating us to veer towards an inductive approach to knowledge discovery. That is, we generate a visualization of the MVD that aims to provide us with non-trivial clues regarding data structure that might lead hypothesis generation [8].

The setting of this exploratory visualization process is as follows. We first consider the classification of a class C GPCR sequence database into each of the seven characteristic subtypes and focus on misclassified cases. Secondly, the same sequences are visualized using a nonlinear dimensionality reduction (NLDR) technique, namely Generative Topographic Mapping (GTM [9]). This technique has been applied with success to many problems in biomedicine and bioinformatics [5, 10–13]. The misclassified cases are then visually isolated and characterized against the rest of the data and, particularly, against the rest of cases of their own subtype. This should help us to differentiate between cases that are likely to be misclassified due to their similarity to overlapping sequences belonging to other subtypes (that is, borderline cases) from those which are misclassified due to an apparently clear wrong subtype assignment. A further visual characterization of the misclassified cases is carried out using phylogenetic trees, which are a standard tool for sequence analysis in bioinformatics.

This exploratory process should help us to build hypotheses about potential database quality problems and about the extent to which GPCR sequence transformations can retain GPCR subtype discriminability. The reported experiments are necessarily limited in their scope, due to space constraints. They are meant as a proof of concept to demonstrate the feasibility of the proposed method as a tool for the detailed analysis of those GPCRs that are consistently misclassified by standard sequence discrimination methods.

2 Materials

The data set analyzed in this study was extracted from version 11.3.4, as of March 2011, of GPCRDB[3] [14] database system for GPCRs, which divides them into several major families or classes based on the ligand types, functions, and sequence similarities. The data set consists of 1,510 GPCRs sequences belonging to class C, which are further subdivided into 7 subtypes: Metabotropic glutamate ($MGlu$), Calcium sensing, GABA-B, Vomeronasal, Pheromone, Odorant and Taste. They are of particular interest for being the target for new therapies in areas such as pain, anxiety, neurodegenerative disorders and as antispasmodics, but also potentially for the treatment of hyperthyroidism and osteoporosis.

The use of transformations of the unaligned sequences allow us to obtain real-valued data matrices to which standard quantitative methods of analysis can be applied. In this study, the very simple AA composition transformation [15] is used as an example for the proof of concept of the proposed visualization method. This is based on the computation of the frequencies of the 20 AAs for each sequence. As a result, a $N \times 20$ matrix is obtained, where $N = 1,510$.

3 Methods

3.1 Visualization Using Manifold Learning Methods

Many methods for MVD visualization are available to the data analyst. NLDR techniques [16], in particular, have undergone a rapid evolution over the last decade, showing great potential as flexible tools for insightful data visualization. A well-know example is Self-Organizing Maps (SOM, [17]), widely used in bioinformatics and biomedicine.

In this study, we use a probabilistic alternative to SOM called GTM [9]. As a manifold learning method, it models the MVD by "covering" them with a low-dimensional manifold. As a Vector Quantization method, it expresses that manifold, in a similar way as SOM, as a network of cluster centroids or data prototypes that, in the case of GTM, are also the centres of distributions. This way, the GTM can be expressed as a manifold-constrained mixture of distributions.

The GTM provides MVD visualization because the model is expressed as a (nonlinear) mapping from a low-dimensional latent visualization space (2-D in this study) into the observed data space, in the form $y = \varPhi(u)W$, where y is a vector in a D-dimensional data space, $\varPhi$ is a set of M basis functions, u is a point in the visualization space and W is the matrix of adaptive weights w_{md}.

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

The probability distribution for data point x in $X = \{x_1, ..., x_N\}$ with $x \in \Re^{\mathrm{D}}$, generated by a latent point u, is defined as an isotropic Gaussian noise distribution, assuming a single common inverse variance β:

$$p\left(x|u, W, \beta\right) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\frac{\beta}{2}\left\|x - y\left(u, W\right)\right\|^2\right\} \tag{1}$$

Integrating out the latent variables u, we can obtain $p\left(x\right)$ and the corresponding likelihood of the model. Standard maximum likelihood methods can then be used to estimate the optimum values of the adaptive parameters. Details can be found in [9]. As part of the parameter estimation process, the probability of each of the K latent points u_k for the generation of each data point x_n can be explicitly calculated as the *responsibility* r_{kn}:

$$r_{kn} = \mathrm{P}\left(k|x_n, W, \beta\right) = \frac{\exp\left\{-\frac{\beta}{2}\left\|x_n - y_k\right\|^2\right\}}{\sum_{k'=1}^{K} \exp\left\{-\frac{\beta}{2}\left\|x_n - y_{k'}\right\|^2\right\}} \tag{2}$$

For MVD visualization, r_{kn} enables a “soft projection”, also known as *posterior mean projection*, defined as $u_n^{mean} = \sum_{k=1}^{K} r_{kn} u_k$. In our experiments GTM parameters were initialized according to a standard PCA-based procedure [9].

3.2 Hierarchical Sequence Visualization Using Phylogenetic Trees

For proteins, a phylogenetic tree (PT) is a dendogram-like graphical representation of the evolutionary relationship between taxonomic groups which share a set of homologous sequence segments. This evolutionary relationship is a form of hierarchically structured similarity-based grouping process. It can be argued that such graphical representation is by itself, a form of data visualization.

Treevolution[4] [18] is a software developed in Java that integrates the Processing[5] package. This tool supports visual and exploratory analysis of PTs in either Newick or PhyloXML formats as radial dendograms, with high-level user-controlled data interaction. The color-guided highlighting of protein families helps the user to focus on sequence groupings of interest. The PT is created from a multiple sequence alignment obtained with Clustal Omega [19]. The PT and GTM sequence visualization approaches differ from each other; the former adopts a hierarchical clustering approach from aligned versions of the sequences and only reflects relative similarity, whereas the latter does not reflect hierarchy but implicitly, while reflecting similarity in projective form. These approaches, though, nicely complement each other and yield quite consistent results.

4 Experiments and Results

Experiments were performed for all the class C subtypes listed in section 2. Due to space limitations, we exemplify the proposed visualization-based method

[4] `http://vis.usal.es/treevolution`

[5] `http://processing.org`

using only *mGlu*. A total of 16 *mGlu* sequences were misclassified by SVM [6]. For illustration, Table 1 lists those 5 predicted to belong to the *Odorant* subtype.

ID	GPCRs name
39	*a8dz71_danre*
40	*a8dz72_danre*
45	*q5i5d4_9tele*
46	*q5i5c3_9tele*
58	*a7rr90_nemve*

Table 1. List of GPCRDB identifiers for the 5 *mGlu* predicted to be *Odorants*, including an index number.

The complete set of class C GPCRs was then visualized using the *posterior mean projection* of GTM, as displayed in Figure 1, left. As we focus on *mGlu*, the isolated visualization of this subtype is shown in Figure 1, right. Sequences that were correctly classified by SVM are displayed as white star symbols, whereas the 16 misclassified ones are represented with the symbol that corresponds to their *predicted* subtype.

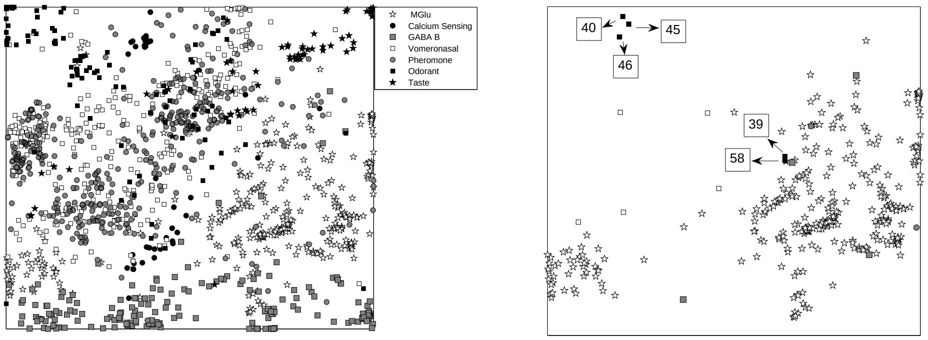

Fig. 1. Left: GTM *posterior mean projection* visualisation of the complete class C GPCR data set. Right: Visualisation of *mGlu* sequences. Cases correctly classified by SVM are displayed as white stars; misclassified ones are represented with the symbols of their predicted subtypes. Cases labeled with ID as in Table 1.

Finally, a phylogenetic tree of the complete set of 1,510 sequences was created and the 5 *mGlu* cases apparently misclassified as *Odorant* were visually isolated, as shown in Figure 2.

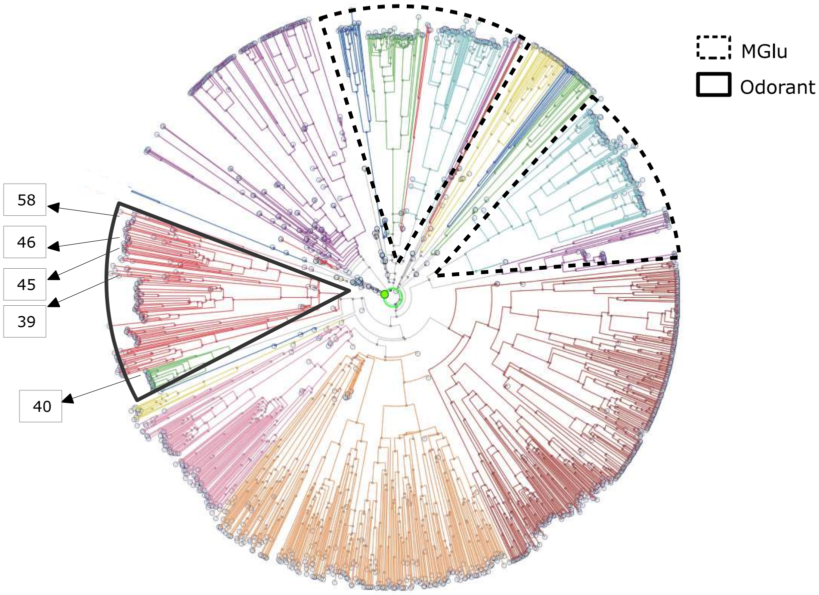

Fig. 2. Treevolution radial PT plot for the 1,510 GPCRs. Each outer branch corresponds to one GPCR sequence. Two separated groups of *mGlu* are identified. The 5 *mGlu* cases misclassified as *Odorant* are singled out within an *Odorant* region. At a given radial distance, tree colors represent families of descendant nodes. For example, the two different colors assigned to *Odorant* provide quantitative evidence of the existence of two subtypes within the family, corresponding to the next subtyping level.

5 Discussion

It is clear from the GTM visualization of the complete set of class C GPCR sequences (Figure 1, left), that there exists a reasonable level of subtype differentiation, but also that some subtypes, such as GABA-B, are more clearly separated from the rest than others, such as *Pheromone* and *Veromonasal*, which strongly overlap. Focusing on the *mGlu* subtype, Figure 1, right, reveals the five sequences that have been misclassified as *Odorant*. They are clearly of an heterogeneous nature: sequences 40, 45 and 46 are clustered together and in a position of the GTM visualization map that fully overlaps the most densely *Odorant*-populated region. Instead, sequences 39 and 58 are quite close to the densest cluster of *mGlu* cases, but in its border and also close to a number of odorant sequences. This comes as no surprise, given the well-documented sequential similarity between certain *Odorant* and *mGlu* receptors [20].

From the previous visualization, it could be hypothesized that these misclassifications might belong to two different types: cases 39 and 58 might be borderline classifications of cases that are close enough to *mGlu* sequences, but not too different to at least some *Odorant*, while cases 40, 45 and 46 would be strong misclassifications that might hint a situation of potential sequence mislabelling. It is not difficult to validate these results using the phylogenetic tree displayed in 2. Despite being labelled as *mGlu* in GPCRDB, all of the five sequences squarely fall in the tree area populated by the *Odorants*, which implies that these sequences are more similar to the latter than to the former subtype. With the support of these visualization-based results, an expert in the field could then inspect these GPCRs under suspicion.

The pair *a8dz71_danre* and *a8dz72_danre*, according to the UniProt[6] database, are uncharacterized proteins, derived from an Ensembl automatic analysis pipeline and should be considered as preliminary data. In fact, Ensembl characterizes them as class C olfactory receptors. According to UniProt and the European Nucleotide Archive[7], *q5i5d4_9tele* and *q5i5c3_9tele* are, in turn, unreviewed putative pheromone receptors CPpr3 and CPpr14. Finally, and also according to UniProt, *a7rr90_nemve* is a predicted protein, where "predicted" qualifies entries without evidence at protein, transcript, or homology levels and which are just one level over "uncertain".

Given that these results are based on the AAC transformation of the GPCR sequences, the AA ratio profiles of each of the misclassified sequences could also be directly inspected by experts to find possible discrepancies with the average profiles of the labeled and predicted subtypes.

6 Conclusions

The classification of class C GPCRs from their transformed unaligned primary sequences seems to have a limiting classification threshold. In this paper, we have proposed a visualization method for the exploration of misclassifications, based on manifold learning models and phylogenetic trees, aimed to detect potential database labelling quality problems. The reported experiments have exemplified, as a proof of concept, the core exploratory data-centered process that should lay the foundations for a full decision support system that, together with prior human expert knowledge, would become a tool for the detailed analysis of those GPCRs that are consistently misclassified by sequence discrimination methods.

Future research should test the method using alternative unaligned transformations of the GPCR sequences. Furthermore, and given that both the GTM and PT have visually revealed substructure within the different class C GPCR subtypes, it should also be investigated at deeper levels of subtyping [3].

References

1. Rask-Andersen, M., Sällman-Almén, M., Schiöth, H.B.: Trends in the Exploitation of Novel Drug Targets. Nat Rev Drug Discov 10, 579–590 (2011)
2. Kniazeff, J., Prézeau, L., Rondard, P., Pin, J.P., Goudet, C.: Dimers and Beyond: The Functional Puzzles of Class C GPCRs. Pharmacol Ther 130, 9–25 (2011)

[6] `http://www.uniprot.org/uniprot/{A8DZ71,A8DZ72}`

[7] `http://www.ebi.ac.uk/ena`

3. Gao, Q.-B., Ye, X.-F., He, J. Classifying G-Protein-Coupled Receptors to the Finest Subtype Level. Biochem Bioph Res Co 439(2), 303–308 (2013)
4. 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, e46,633 (2012)
5. Cruz-Barbosa, R., Vellido, A., Giraldo, J.: Advances in Semi-Supervised Alignment-Free Classification of G-Protein-Coupled Receptors. In: Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO'13), Granada, Spain, pp. 759–766 (2013)
6. König, C., Cruz-Barbosa, R., Alquézar, R. and Vellido, A. SVM-Based Classification of Class C GPCRs from Alignment-Free Physicochemical Transformations of Their Sequences. In A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 336–343, (2013)
7. Rehman, Z.U., Mirza, M.T., Khan, A., Xhaard, H. Predicting g-protein-coupled receptors families using different physiochemical properties and pseudo amino acid composition. Method enzymol 522, 61–79 (2012)
8. Vellido, A., Martín, J.D., Rossi, F., Lisboa, P.J.G.: Seeing is Believing: The Importance of Visualization in Real-World Machine Learning Applications. In: Proceedings of the 19^{th} European Symposiun on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2011), pp.219–226, d-side pub. (2011)
9. Bishop, C.M., Svensén, M., Williams, C.K.I.: GTM: The Generative Topographic Mapping. Neural Comput 10, 215–234 (1998)
10. Vellido, A., Romero, E., Julià-Sapé, M., Majós, C., Moreno-Torres, À., Arús, C.: Robust Discrimination of Glioblastomas from Metastatic Brain Tumors on the Basis of Single-Voxel Proton MRS. NMR Biomed 25(6), 819–828 (2012)
11. Cárdenas, M.I., Vellido, A., Olier, I., Rovira, X., Giraldo, J.: Complementing Kernel-Based Visualization of Protein Sequences with Their Phylogenetic Tree, LNCS/LNBI 7548, pp. 136–149 (2012)
12. Mumtaz, S., Nabney, I. T., Flower, D.: Novel Visualization Methods for Protein Data. In: 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 198–205 (2012)
13. Vellido, A., Cárdenas, M.I., Olier, I., Rovira, X., Giraldo, J.: A Probabilistic Approach to the Visual Exploration of G Protein-Coupled Receptor Sequences. In: Proceedings of the 19^{th} European Symposiun on Artificial Neural Networks , Computational Intelligence and Machine Learning (ESANN 2011), pp. 233–238, d-side publishing, Belgium (2011)
14. Vroling, B., Sanders, M., Baakman, C., Borrmann, A., Verhoeven, S., Klomp, J., Oliveira, L., de Vlieg, J., Vriend, G.: GPCRDB: information system for G protein-coupled receptors. Nucleic Acids Res 39(suppl 1), D309-D319 (2011)
15. 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)
16. Lee, J.A., Verleysen, M. Nonlinear Dimensionality Reduction. Springer (2007)
17. Kohonen, T.: Self-Organizing Maps (3^{rd} ed.). Springer (2001)
18. Santamaría, R., Therón, R. Treevolution: visual analysis of phylogenetic trees. Bioinformatics, 25(15), 1970-1971 (2009)
19. Sievers, F. et al.: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539 (2011)
20. Kuang, D., Yao, Y., Wang, M., Pattabiraman, N., Kotra, L.P., Hampson, D.R. Molecular Similarities in the Ligand Binding Pockets of an Odorant Receptor and the Metabotropic Glutamate Receptors. J Biol Chem 278(43), 42551–42559 (2003)