GENOMICS AND COMPUTATIONAL BIOLOGY Vol. 1, No. 1 (2015): e19

Research Article

Visual Characterization of Misclassified Class C GPCRs through Manifold-based Machine Learning Methods

Martha I. Cárdenas[1,2*], Alfredo Vellido [1,3], Caroline König[1], René Alquezar[1] and Jesús Giraldo[2]

[1]Departament de Ciències de la Computació, Universitat Politècnica de Catalunya, 08034, Barcelona, Spain
[2]Institut de Neurociències, Unitat de Bioestadística, Universitat Autònoma de Barcelona, 08193, Bellaterra, Barcelona, Spain
[3]Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Cerdanyola del Vallès, Barcelona, Spain

*To whom correspondence should be addressed: mcardenas@cs.upc.edu

Received: 2015-02-15; Accepted: 2015-07-13; Published: 2015-09-18

ABSTRACT

G-protein-coupled receptors are cell membrane proteins of great interest in biology and pharmacology. Previous analysis of Class C of these receptors has revealed the existence of an upper boundary on the accuracy that can be achieved in the classification of their standard subtypes from the unaligned transformation of their primary sequences. To further investigate this apparent boundary, the focus of the analysis in this paper is placed on receptor sequences that were previously misclassified using supervised learning methods. In our experiments, these sequences are visualized using a nonlinear dimensionality reduction technique and phylogenetic trees. They are subsequently characterized against the rest of the data and, particularly, against the rest of cases of their own subtype. This exploratory visualization should help us 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; pharmacoproteomics

INTRODUCTION

G-protein-coupled receptors (GPCRs) were the subject matter of the 2012 Nobel Prize in Chemistry [1]. They comprise a large superfamily of membrane proteins that play an important role in cell communication and currently constitute a major target for drug discovery. For these reasons, they are intensely investigated in the field of pharmacoproteomics [2].

The GPCR superfamily can be classified into four main classes, namely A, B, C, and F (Frizzled) according to their similarity [3]. Class C GPCRs in particular, which are the focus of our research, show a remarkable diversity in terms of both structure and functional roles. Thus, the correct discrimination of class C GPCRs according to subtypes is a challenging classification problem which constitutes the starting point for our study [4, 5].

In addition to the seven transmembrane (7TM) domain, which is common to all GPCRs, class C GPCRs are characterized by bearing a large extracellular domain, the Venus flytrap (VFT), in which orthosteric ligands bind. This structural complexity has prevented the crystallization of full-length class C GPCRs, and was not till 2014 that the 7TM domains of two members of this family were crystallized [6, 7]. Because of this, the investigation of class C GPCR structure and function on the basis of their primary amino acid (AA) sequences is of special relevance.

The unaligned symbolic sequences do not yield themselves easily to direct quantitative analysis, but many different primary sequence transformation techniques are available to overcome this limitation. In this study, we use transformed alignment-free full sequences to limit information loss.

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 and supervised classification of class C GPCRs [8, 9] with this type of transformation showed that overall accuracy (the ratio of correctly classified sequences) reaches an upper bound in the area of 90% that it is not significantly increased when more sophisticated physico-chemical transformations of the sequences are applied.

To investigate this apparent classification boundary, 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 in our work as an exploratory Data Mining tool, facilitating the analyst to veer towards an inductive approach to knowledge discovery. That is, we generate a visualization of the MVD that aims to provide the analyst with non-trivial clues regarding data structure that might lead to hypothesis generation [10, 11].

A further and complementary visual grouping characterization of the class C GPCRs is carried out using phylogenetic trees (PTs), which are a standard

bioinformatics tool for protein analysis from aligned sequences.

The setting of this exploratory visualization process is as follows. We first consider the classification of a class C GPCR sequence dataset into each of its seven characteristic subtypes and proceed to single out misclassified cases. Secondly, the same sequence dataset is visualized using a nonlinear dimensionality reduction (NLDR) technique, namely Generative Topographic Mapping (GTM [12]). This technique has been applied with success to many problems in biomedicine and bioinformatics [8, 13-15].

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 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. The latter can also be understood as part of a *label noise* problem [16], in which the possibility of wrong class labeling is accepted and addressed in different ways.

This exploratory process should help the analyst to build hypotheses about potential database quality problems (in the form of potentially inadequate subtype labels) and about the extent to which GPCR sequence transformations can retain GPCR subtype discriminability. The reported experiments are meant to be 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.

MATERIALS AND METHODS

Class C GPCR data

The data set analyzed in this study was extracted from version 11.3.4, as of March 2011, of GPCRDB[a] [17], a database information system for GPCRs that includes sequential data. The system divides GPCRs into several major families or classes based on the ligand types, functions, and sequence similarities.

The analyzed dataset consists of 1,510 GPCRs sequences that belong to class C. This class is 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.

Class C sequences in our dataset are, in turn, distributed into 7 subtypes: 351 cases of metabotropic glutamate, 48 calcium sensing, 208 $GABA_B$, 344 vomeronasal, 392 pheromone, 102 odorant and 65 taste (see Table 1 for details and the abbreviations that will be used throughout the text).

The lengths of these sequences varied from 250 to 1,995 AA, a wide range that provides further justification for the use of alignment-free sequence transformation strategies. The varying lengths of the receptors in the analyzed data do not seem to have an important effect on their assignment to the different subtypes by Support Vector Machine (SVM) classifiers that are the starting point of this study; this conclusion is supported by the results compiled in Supplementary File 1.

Subtype ID	Subtype Description	Number of sequences
mGlu	Metabotropic glutamate	351
CaS	Calcium sensing	48
GABA-B	$GABA_B$	208
VN	Vomeronasal	344
Ph	Pheromone	392
Od	Odorant	102
Ta	Taste	65

Table 1. Class C GPCRs dataset. The 1,510 sequences are structured into 7 subtypes. For each subtype, this table displays the abbreviation identifier (ID), the description and the corresponding number of sequences.

The use of transformations of the unaligned sequences allows us to obtain real-valued data matrices to which standard quantitative methods of analysis can be applied. In the experiments reported in this study, the very simple AA composition (AAC) transformation [18] is used as an example for the proof of concept of the proposed visualization-based method. In this transformation, the frequencies of the 20 AAs are computed for each sequence. As a result, a $N \times 20$ matrix is obtained, where $N = 1{,}510$.

Visualization Using Manifold Learning Methods

Many methods for MVD visualization are available to the data analyst. NLDR techniques, in particular, have undergone a rapid evolution over the last decade, showing great potential as flexible tools for insightful data visualization [19].

Self-Organizing Maps (SOM, [20]), widely used in bioinformatics and biomedicine, are a well-known example of these. In the current paper, we use an alternative to SOM with sound probabilistic foundations, called GTM [12]. As a manifold learning method, it models the MVD by "covering" them with a low-dimensional manifold. As a Vector Quantization one, it expresses that manifold, in a similar way as SOM, as a connected network of cluster centroids or data prototypes that, in the case of the standard GTM, are also the centres of Gaussian distributions. This way, the GTM can be expressed as a manifold-constrained mixture of distributions. The probabilistic definition of GTM makes the optimization of the model possible within a Maximum Likelihood approach, which ensures the convergence of the model training error towards a minimum, something for which there is no theoretical guarantee in the case of SOM. The definition of GTM using Bayesian probability theory principles allows it to be extended in a principled manner. These extensions of the original model include, amongst others: automatic regularization (to avoid the data overfitting), variational reformulations, multivariate time series modelling as manifold-constrained Hidden Markov Models, etc.

Being, as previously mentioned, a constrained mixture of distributions model, GTM also allows the choice of different suitable probability distributions (as basis functions) for different types of data.

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 $y = \phi(u)W$, where y is a vector in a D-dimensional data space, ϕ is a set of M basis functions, u is a point in the visualization space and $W$is the matrix of weights w_{md}, adaptively optimized as part of the model learning process.

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

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

Integrating the latent variables u out, we obtain $p(x)$ 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 [12]. 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}:

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

For MVD visualization, r_{kn} enables a "soft projection", also known as *posterior mean projection*, defined as ${u_n}^{mean} = \sum_{k=1}^{K} r_{kn} u_k$. This is a further advantage over the "crisp" assignment of data instances to clusters created by SOM, for which a probability of membership of each instance to each cluster cannot be calculated.

In our experiments, the GTM parameters were initialized according to a standard Principal Component Analysis (PCA)-based procedure [12].

Phylogenetic Trees

A PT of a group of protein sequences is a dendrogram-like graphical representation of the evolutionary relationship between taxonomic groups which share a set of homologous sequence segments. This evolutionary relationship can be represented through a hierarchically structured similarity-based grouping process. Such process thus yields a form of sequence visualization that can complement those provided by the GTM-based methods.

Treevolution[b] [21] is a software tool developed in Java that integrates the Processing© package[c]. It was used in our experiments to create the PTs from multiple sequence alignments (MSA) obtained with Clustal Omega [22]. This tool supports visual and exploratory analysis of PTs in either Newick or PhyloXML formats as radial dendrograms, with high-level user-controlled data interaction. The color-based handling of protein sub-groups helps the user to focus on relevant sequence groupings.

The PT and GTM sequence visualization approaches differ in several ways; the former uses hierarchical clustering from aligned versions of the sequences and only reflects their relative similarity, whereas the latter only reflects hierarchy implicitly, but reflects similarity explicitly as inter-point distances in the projective space. Despite their differences, these approaches, though, nicely complement each other and yield quite consistent results.

RESULTS

Experiments using the proposed visualization-based methods were performed for the class C GPCR data set described in Table 1.

A batch of previous supervised classification experiments using SVMs were the starting point for these [9]. Such experiments involved an iterative 5 cross-validation (CV) process, splitting the dataset into 5 randomly stratified folds where 4 folds were used for the construction of the model and the remaining one to evaluate the classification results. This process was repeated 100 times and in these experiments, different sequences from each of the seven GPCR subtypes were consistently misclassified (see summary information in Table 2).

Subtype ID	Number of misclassified sequences
mGlu	16
CaS	5
GABA-B	8
VN	46
Ph	48
Od	35
Ta	5

Table 2. Number of class C misclassified sequences, listed by subtype.

Sequence ID	Predicted subtype	Sequence name
39	Od	*a8dz71_danre*
40	Od	*a8dz72_danre*
45	Od	*q5i5d4_9tele*
46	Od	*q5i5c3_9tele*
58	Od	*a7rr90_nemve*
60	GABA-B	*a7rrr9_nemve*
105	GABA-B	*d1lx28_sacko*
142	GABA-B	*XP_002735016*
206	GABA-B	*XP_968952*
59	VN	*a7rsa2_nemve*
66	VN	*b3rud7_triad*
140	VN	*XP_002161343*
141	VN	*XP_002732197*
244	VN	*a7s4n3_nemve*
135	Ph	*a7ria2_nemve*
259	Ph	*q62916_rat*
Total mGlu		16 sequences

Table 3. Misclassified mGlu sequences. List of the 16 misclassified mGlu, including their GPCRDB identifier (ID), their class as predicted by SVM and their sequence name.

In previous preliminary research for the current paper, the 16 misclassified mGlu transformed sequences were analyzed in some detail [23]. We now extend these experiments to the rest of GPCR subtypes and their most consistently 163 misclassified sequences.

Sequence ID	Predicted subtype	Sequence name
372	mGlu	*XP_002123664*
352	VN	*q5i5c8_9tele*
353	VN	*a8e7u1_danre*
370	Ph	*XP_001515899*
399	Ph	*XP_002740613*
Total CaS		5 sequences

Table 4. Misclassified CaS sequences. List of the 5 misclassified CaS, including their GPCRDB identifier (ID), their class as predicted by SVM and their sequence name.

ID	Predicted class	GPCRs name
521	mGlu	*XP_002123664*
530	mGlu	*q5i5c8_9tele*
542	VN	*a8e7u1_danre*
414	mGlu	*a7rpp5_nemve*
494	mGlu	*b3rj55_triad*
486	mGlu	*b3rit4_triad*
475	mGlu	*a7s6r9_nemve*
535	mGlu	*XP_002738008*
Total GABA-B		8 sequences

Table 5. Misclassified GABA-B sequences. List of the 8 misclassified GABA-B, including their GPCRDB identifier (ID), their class as predicted by SVM and their sequence name.

Sequence ID	Predicted subtype	GPCRs name
1450	GABA-B	*q4rx46_tetng*
1451	VN	*q4rx45_tetng*
1462	VN	*a4phq8_danre*
1471	Ph	*XP_425740*
1505	Ph	*q4s833_tetng*
Total Ta		5 sequences

Table 6. Misclassified Ta sequences. List of the 5 misclassified Ta, including their GPCRDB identifier (ID), their class as predicted by SVM and their sequence name.

Subtype	Predicted subtype	Number of misclassificatons
VN	mGlu	7
VN	CaS	2
VN	Ph	30
VN	Od	7
Ph	mGlu	19
Ph	GABA-B	4
Ph	VN	22
Ph	Od	3
Od	mGlu	4
Od	VN	14
Od	Ph	17

Table 7. Misclassified VN, Ph and Od sequences. Summary list of the largest groups of misclassifications.

Tables 3 to 6 list in detail all the misclassified sequences from mGlu, CaS, GABA-B and Ta subtypes. For the sake of brevity, the characteristics of the far more abundant Vn, Ph and Od subtypes misclassifications are summarily reported in Table 7 and reported in full as Supplementary File 2.

GTM posterior mean projection visualization

The complete class C GPCR dataset, including 1,510 sequences, was then visualized using the *posterior mean projection* of GTM, as described in previous sections. This global GTM visualization map is displayed in Figure 1. Note that the axes in the representation space have no units because each of them represents one of the dimensions of the latent space of the GTM model.

Each of the subtypes is then represented in isolation in the GTM maps of Figures 2 to 8. In each of these maps, the misclassified sequences are individually identified using the sequence ID.

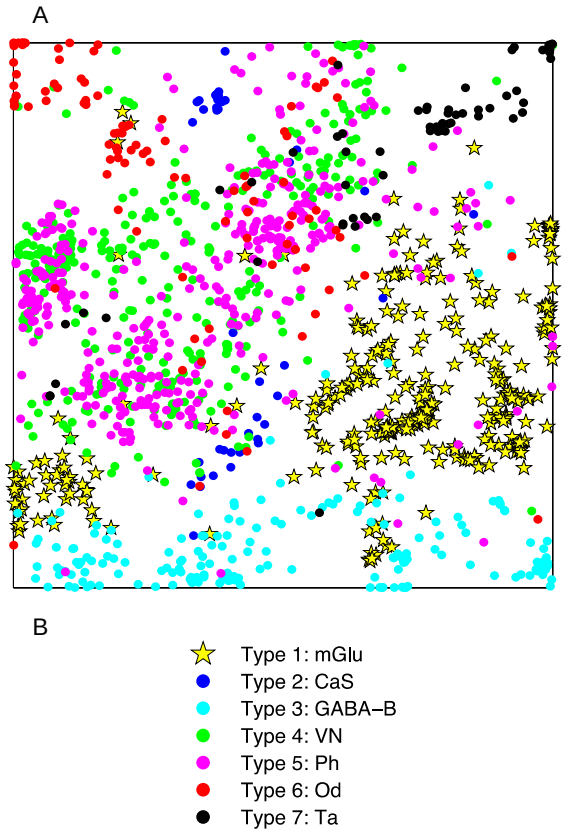

Figure 1. Dataset GTM posterior mean projection (A) and list of corresponding labels (B). Visualization of all 1,510 sequences. Each color corresponds to a GPCR class C subtype.

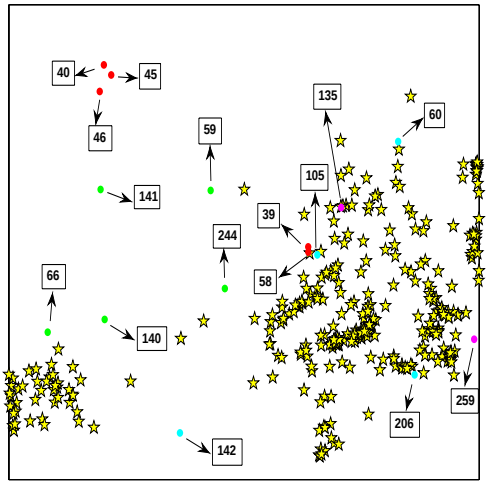

Figure 2. mGlu GTM posterior mean projection. Visualization of mGlu sequences. Cases incorrectly classified by SVM are represented with the colors of their predicted subtypes. Cases labeled with their ID from Table 3.

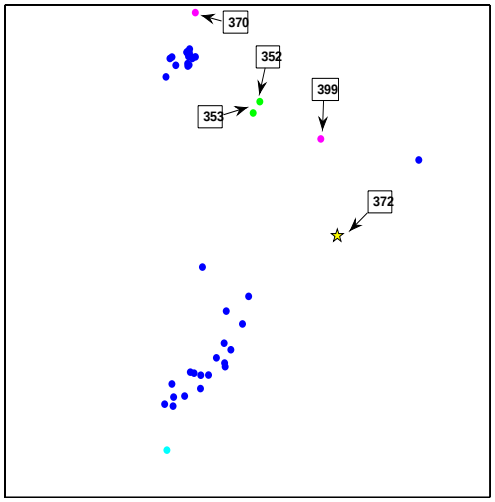

Figure 3. CaS GTM posterior mean projection. Visualization of CaS sequences. Representation as in Figure 2.Cases labeled with their ID from Table 4.

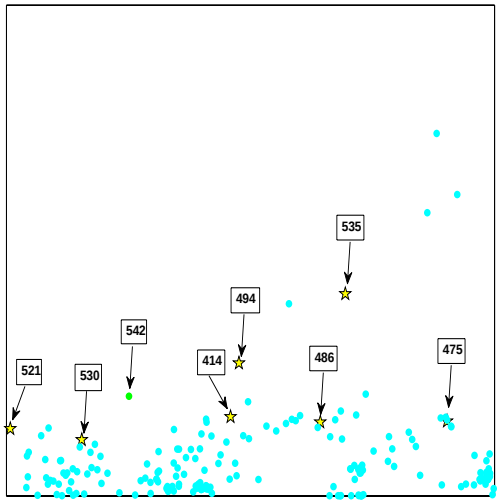

Figure 4. GABA-B GTM posterior mean projection. Visualization of GABA-B sequences. Representation as in Figure 2.Cases labeled with their ID from Table 5.

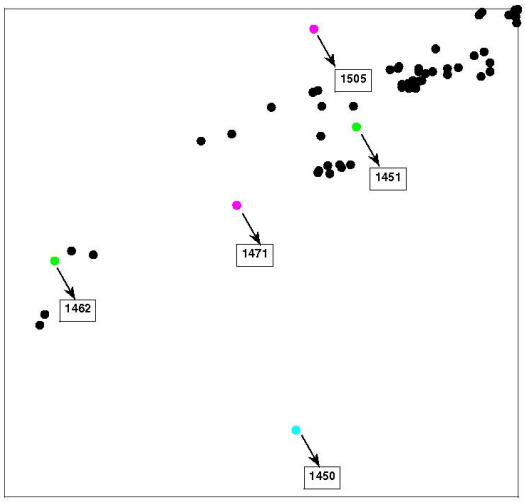

Figure 5. Ta GTM posterior mean projection. Visualization of Ta sequences. Representation as in Figure 2. Cases labeled with their ID from Table 6.

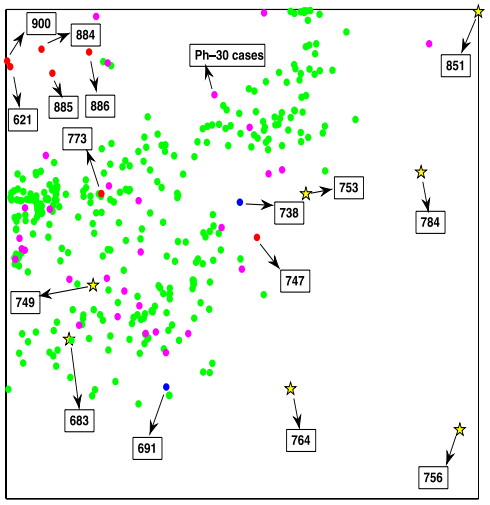

Figure 6. VN GTM posterior mean projection. Visualization of VN sequences. Representation as in Figure 2. Note that the 30 Ph misclassified cases are not individually labeled.

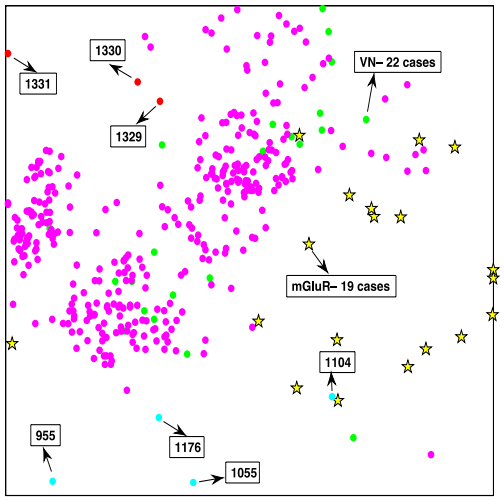

Figure 7. Ph GTM posterior mean projection. Visualization of Ph sequences. Representation as in Figure 2. Note that the 22 VN and 19 mGlu misclassified cases are not individually labeled.

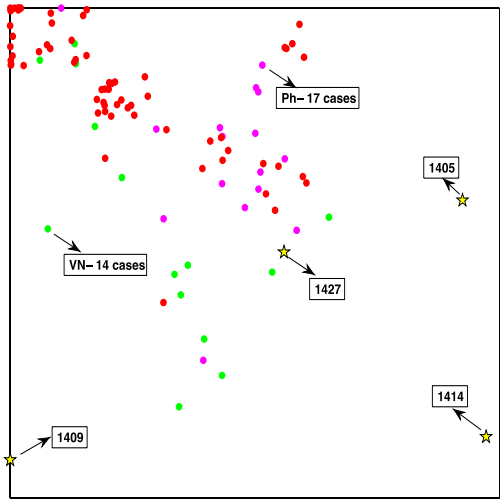

Figure 8. Od GTM posterior mean projection. Visualization of Od sequences. Representation as in Figure 2. Note that the 14 VN and 17 Ph misclassified cases are not individually labeled.

Treevolution radial PT

Finally, a phylogenetic tree of the complete set of 1,510 sequences was created using Treevolution software. It is shown in Figure 9 and will be used to highlight the misclassifications listed in the previous section.

The Radial PT supports interactive exploration according to the hierarchical structure it provides. At a given radial distance, different colors represent the same family of descendant nodes in the tree.

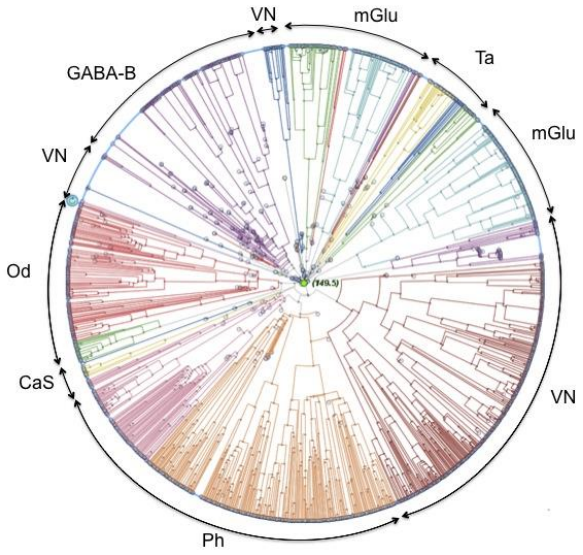

Figure 9. Treevolution radial PT plot of the 1,510 GPCRs. Each branch corresponds to one GPCR sequence. Two separated mGlu sections can be identified, as well as three consecutive CaS sections; a single GABA-B section; three separate VN ones; two consecutive groups of Ph; two of Od and three consecutive groups of Ta. At a given radial distance, the 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 at a deeper level in the hierarchy.

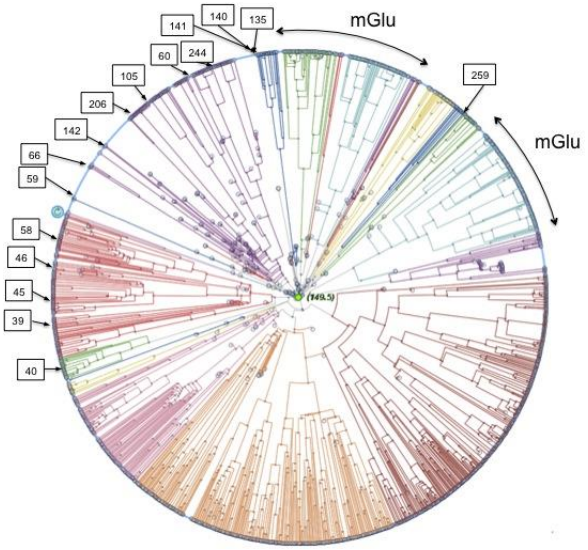

Figure 10. Radial PT plot for mGlu misclassified cases.

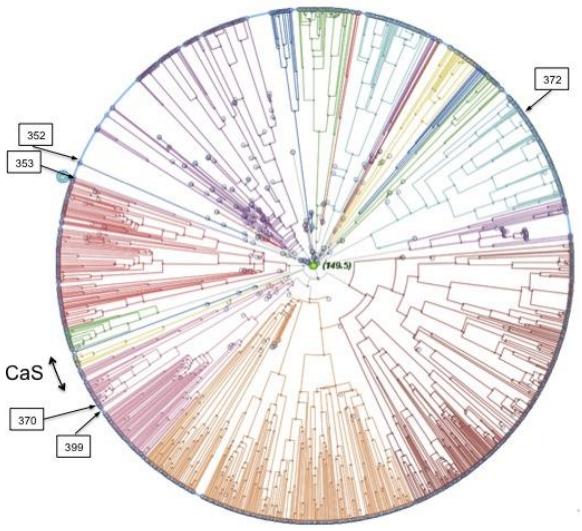

Figure 11. Radial PT plot for CaS misclassified cases.

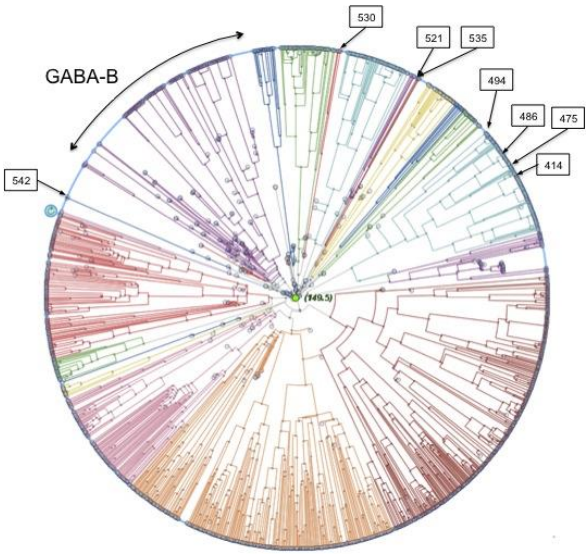

Figure 12. Radial PT plot for GABA-B misclassified cases.

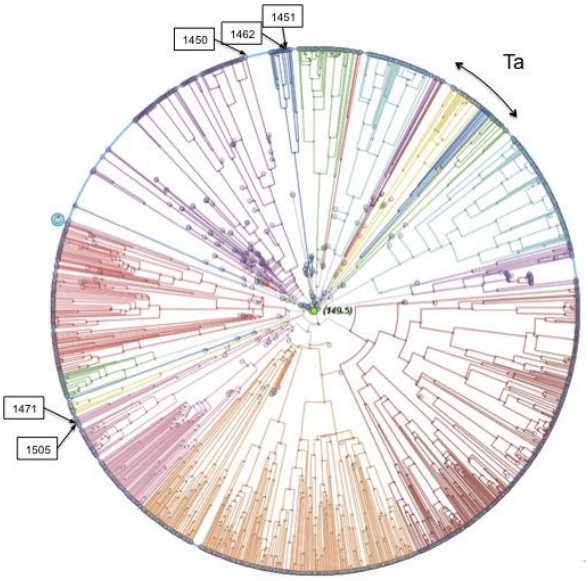

Figure 13. Radial PT plot for Ta misclassified cases.

DISCUSSION

It is clear from the GTM visualization of the complete set of transformed class C GPCR sequences (Figure 1), 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 Vomeronasal, which strongly overlap. The overlapping (or its lack) of subtype data projections in the GTM map should be a solid indication of subtype discriminability (or lack of it).

Focusing first on the mGlu subtype, Figure 2 reveals quite clear patterns of misclassification. See, for instance, sequences 40, 45 and 46. They are clustered together and in a position of the GTM visualization map that fully overlaps the most densely Odorant-populated region (as seen in Figure 8). These cases could be understood as neat, strong misclassifications and, therefore, worth investigating as potential cases of label noise. The same could be said of, at least, sequences 59, 140, 141 and 142, which have been misclassified as either GABA-B or VN due to the fact that they are clearly positioned in their corresponding regions.

Instead, sequences 39 and 58, misclassified as Od, are located quite close to the densest cluster of mGlu cases, but nearby its boundaries and also close to a number of actual Od sequences. This comes as no surprise, given the well-documented sequential similarity between certain Odorant and mGlu receptors [24]. These cases might therefore be considered as borderline misclassifications of sequences that are close enough to mGlu, but not too different to at least some Od.

A similar distinction between strong and borderline misclassifications can be found for the remaining class C subtypes. In the case of CaS, which shows two neatly differentiated subgroups that indicate (as in the case of mGlu) further levels of sub-structure, all five of the misclassified sequences (as either mGlu, VN, or Ph) seem to belong to the strong misclassification category, again meriting further inspection as potential cases of label noise.

The case of Ta is almost the opposite: although, again, a clear two-subgroup structure can be found, it could be argued that all but one of the five misclassified sequences (as VN, or Ph) are, in fact, borderline cases. Instead, case 1450 is strongly misclassified as a GABA-B, falling squarely within the domain area of this subtype.

The situation for GABA-B is not too dissimilar. Most misclassifications are borderline cases that get confused as mGlu given the partial overlap of both subtypes. The only exception might be case 535, deep within the central mGlu map domain.

The remaining subtypes, namely VN, Ph and Od, experiment a very strong level of overlapping with other subtypes and, as result, borderline misclassifications abound. In the case of Ph, there is a sizeable number of cases strongly misclassified as mGlu and a few as GABA-B and Od. For VN, instead, only a few cases are strongly misclassified as mGlu, but a few more as Od. Finally, Od, again a subtype evidencing further sub-structure, has quite a few cases strongly misclassified as VN and Ph.

With the support of these visualization-based results, an expert in the field could smoothly move from exploratory visualization to the detailed inspection of the strongly misclassified class C GPCRs as potential suspects of mislabeling in a case of label noise.

For the mGlu cases strongly misclassified as Od (see Table 3), for instance, the pair a8dz71_danre and a8dz72_danre, according to the UniProt[d] 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[e], 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".

For the CaS cases, q5i5c8_9tele, misclassified as VN is, according to UniProt, Putative pheromone receptor CPpr9 and its status is "unreviewed" (not manually annotated and reviewed by UniProt curators); a8e7u1_danre (again misclassified as VN) is both "unreviewed" and "uncharacterized". XP_001515899 and XP_002740613 are misclassified as pheromones: the former has been predicted to be similar to a calcium-sensing receptor[f], whereas the latter was "removed as a result of standard genome annotation processing" from NCBI[g] Finally, XP_002123664, misclassified as an mGlu, was also "removed as a result of standard genome annotation processing" from NCBI[h], despite being previously predicted to be similar to a calcium-sensing receptor.

The Taste q4rx46_tetng, strongly misclassified as GABA-B, is identified by UniProt as the *unreviewed Chromosome 11 SCAF14979, whole genome shotgun sequence*.

The GABA-B XP_002738008, misclassified as mGlu, is, interestingly, predicted in NCBI[i] to be an extracellular calcium-sensing receptor.

The remaining three subtypes have a strongly overlapping behavior that suggests that the current AAC transformation does not suffice to discriminate them properly and include too many strong misclassifications to individually discuss in detail. Nevertheless the proposed visualization-based method would provide the expert with guidance to inspect any of these cases as required.

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.

Figure 9 displays the complete radial PT for the 1,510 sequences and outlines the main domains of all seven class C subtypes in its external border. Even though the original sequence transformations have very little in common with those used in the GTM-based visualization (bear in mind that the PT is built from aligned sequences), the misclassification results reported in detail in Figures 10 to 13 for, in turn,

subtypes mGlu, CaS, GABA-B and Ta are quite consistent with those shown in GTM Figures 2 to 5. Although the results are similar for VN, Ph and Od, they are again not included here due to the large amount of misclassified cases involved.

Each individual misclassified sequence is identified with its corresponding ID. In Figure 10, for example, where mGlu sequences are highlighted, the five sequences predicted as Odorants squarely fall in the tree area populated by this subtype, which implies that these sequences are more similar to the latter than to the mGlu subtype to which they are assumed to belong according to their label in GPCRDB. Similarly, the four GABA-B, five VN and two Ph sequences displayed in Figure 10 are located in the corresponding areas of their predicted subtypes.

The results visualized in Figures 11, 12 and 13 CaS, GABA-B and Ta, respectively, fully agree with those discussed for mGlu, with misclassified sequences located in the domains of the predicted subtypes, instead of in the domains of their database label.

Note that it is far more difficult to distinguish between borderline and strong misclassifications in the radial PTs due to the intrinsic symmetry of their branches.

CONCLUSIONS

In this paper, we have analyzed class C GPCR full unaligned primary sequences transformed according to a simple amino acid frequency method.

Prior research had revealed a limit on the ability to discriminate these transformed sequences into their seven known subtypes, prompting suspicion that, at least partially, this could be caused by sequence mislabeling, a type of label noise [16, 23].

We have proposed a method to investigate misclassified class C GPCRs that is based on NLDR, manifold-based visualization, complemented by the use of PTs. This method has revealed that, for each of the analyzed subtypes, misclassified sequences are either *borderline* cases, whose label might have been incorrectly predicted due to lack of sensitivity of the classifier, or *strong* misclassifications that are truly similar to sequences belonging to other subtypes.

The latter are of special interest for database quality assessment purposes and our discussion of the reported results has shown that many of the cases singled out for further inspection were in fact unresolved or unclear subtype assignments according to main protein database repositories such as UniProtKB/Swiss-Prot and GenBank-NCBI.

At the heart of this investigation on the limitations of classifiers in the characterization of labeled class C GPCRs, lies the fact that proteins in curated databases are often assigned to families according to data-based models. An example of this is the comprehensive Pfam database [25], built using hidden Markov models and MSA. This is, indeed, a perfectly adequate approach, but even in Pfam-defined families, there are two levels of quality (A and B), where the A entries are derived from the underlying sequence database built from the most recent release of UniProtKB and the B entries are un-annotated and automatically generated, built from sequence clusters not covered by Pfam-A entries. We reckon that the lack of a gold-standard for class C GPCR labelling is what makes our investigation on potential labelling inconsistencies relevant. In addition, it could be particularly useful given the absence of 3D crystal structures for the full sequences of these receptors.

In conclusion, the reported experiments provide a proof of concept for a support method for experts in GPCR (and proteins in general) database quality control and curation.

ACKNOWLEDGEMENTS

This research was partially funded by MINECO TIN2012-31377 and ERA-NET NEURON PCIN-2013-018-C03-02, as well as Fundació La Marató de TV3 110230 research projects.

AUTHOR CONTRIBUTIONS

MC and CK carried out, in turn, the unsupervised and supervised experiments. AV and RA provided machine learning expertise, while JG provided expert support in the biological elements of the research. All authors contributed to the experimental design, as well as to write, revise and approve the manuscript.

CONFLICT OF INTEREST DECLARATION

The authors declare no conflict of interest.

SUPPLEMENTARY DATA

High resolution files of the main figures and supplementary items listed below are available for download at Genomics and Computational Biology online.

Supplementary File 1. The effect of sequence size on class C GPCR subtype classification; File: GCB_Supplementary File 1.docx. The experiments reported in this paper use the AAC sequence transformation and, therefore, the analyzed data consist of vectors of 1-gram frequencies of the same length for every sequence, regardless its original length. We might expect this transformation to limit undesired effects due to the differences in length of the original sequences on the classification of the sequences using SVMs (the starting point of our study). This supplementary file provides some evidence to support this expectation. For that, we show, next to each other, a histogram of the lengths of the complete data set (1,510 sequences) and a histogram of the lengths of the 163 SVM-misclassified sequences.

Supplementary File 2. Annex with tables of sequences misclassified in the SVM-based procedure that were not included in the main text; File: GCB_Supplementary File 2.docx. In this supplementary file, we provide the interested reader with the complete tables of sequences misclassified in the SVM-based procedure that were not included in the main text for the sake of brevity. They belong to class C GPCR subtypes Vomeronasal, Pheromone and Odorant.

ABBREVIATIONS

7TM: 7-TransMembrane

AA: Amino Acid

AAC: Amino Acid Composition

GPCR: G Protein-Coupled Receptor

GTM: Generative Topographic Mapping

MSA: Multiple Sequence Alignment

MVD: MultiVariate Data

NLDR: Non-Linear Dimensionality Reduction

PCA: Principal Component Analysis

PT: Phylogenetic Tree

SOM: Self-Organizing Maps

SVM: Support Vector Machines

REFERENCES

1. The Nobel Prize in Chemistry 2012. Nobelprize.org. Nobel Media AB 2014. Web. 18 Jan 2015 <http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2012/>.
2. Overington JP, Al-Lazikani B, Hopkins Al: **How many drugs are there?** *Nat Rev Drug Discov* 2006*;* **5(12):**993-996. doi:10.1038/nrd2199
3. Lagerström MC, Schiöth HB: **Structural diversity of G protein-coupled receptors and significance for drug discovery.** *Nat Rev Drug Discov* 2008*;* **27:**339-357. doi:10.1038/nrd2518
4. Venkatakrishnan AJ, Flock T, Prado DE, Oates ME, Gough J, Madan Babu M: **Structured and disordered facets of the GPCR fold.** *Curr Opinion Struct Biol* 2014; **27:**129-137. doi:10.1016/j.sbi.2014.08.002
5. Gao QB, Ye XF, He J: **Classifying G-protein-coupled receptors to the finest subtype level.** *Biochem Biophys Res Commun* 2013; **439:**303-308. doi:10.1016/j.bbrc.2013.08.023
6. Wu H, Wang C, Gregory KJ, Han GW, Cho HP, Xia Y, Niswender CM, Katritch V, Meiler J, Cherezov V, Conn PJ, Stevens RC: **Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator.** *Science* 2014; **344(6179):** 58-64. doi:10.1126/science.1249489
7. Doré AS, Okrasa K, Patel JC, Serrano-Vega M, Bennett K, Cooke RM, Errey JC, Jazayeri A, Khan S, Tehan B, Weir M, Wiggin GR and Marshall FH: **Structure of a class C GPCR metabotropic glutamate receptor 5 transmembrane domain.** *Nature* 2014; **551:** 557-562. doi:10.1038/nature13396
8. Cruz-Barbosa R, Vellido A, and Giraldo J: **The Influence of Alignment-Free Sequence Representations on the Semi-Supervised Classification of Class C G Protein-Coupled Receptors**. *Med Biol Eng Comput.* 2015; **53**(2): 137-149. doi:10.1007/s11517-014-1218-y
9. 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 *New Trends in Image Analysis and Processing, ICIAP*. Edited by Petrosino A. *et al.*, *LNCS Vol.8158*, Springer; 2013; 336-343. doi:10.1007/978-3-642-41190-8_36
10. Keim DA, Mansmann F, Schneidewind J, Thomas J, and Ziegler H: **Visual Analytics: Scope and Challenges**. In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics. 2008*. doi:10.1007/978-3-540-71080-6_6
11. Fry BJ: **Computational Information Design.** PhD dissertation, Massachusetts Institute of Technology. 2004.
12. Bishop CM, Svensén M, Williams CKI: **GTM: The Generative Topographic Mapping**. *Neural Comput*. 1998; **10:**215-234. doi:10.1162/089976698300017953
13. Cárdenas MI, Vellido A, Olier I, Rovira X, Giraldo J: **Complementing Kernel-Based Visualization of Protein Sequences with Their Phylogenetic Tree**. In *Proceedings of the 8th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatics CIBB 2011.*Edited by Biganzoli E, Vellido A, Ambrogi, F and Tagliaferri, R. LNCS/LNBI Vol.7548, Springer. 2012;136-149. doi:10.1007/978-3-642-35686-5_12
14. Mumtaz S, Nabney IT, Flower D: **Novel Visualization Methods for Protein Data.** In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB.* 2012;198-205. doi:10.1109/CIBCB.2012.6217231
15. 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* 2012; **25**(6):819-828. doi:10.1002/nbm.1797
16. Frénay B, Verleysen M: **Classification in the presence of label noise: a survey.** *IEEE T Neural Networ* 2014; **25**(5):845-869. doi:10.1109/TNNLS.2013.2292894
17. Vroling B, Sanders M, Baakman C, Borrmann A, Verhoeven S, Klomp J, Oliveira L, de Vlieg J, and Vriend G: **GPCRDB: information system for G protein-coupled receptors.** *Nucleic Acids Res* 2011; **suppl 1**(39):309-319. doi:10.1093/nar/gkq1009
18. Sandberg M, Eriksson L, Jonsson J, Sjöström M, and Wold S: **New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids**. *J Med Chem* 1998; **41**:2481-2491. doi:10.1021/jm9700575
19. Lee JA, Verleysen M: ***Nonlinear Dimensionality Reduction.*** Springer; 2007.
20. Kohonen T: ***Self-Organizing Maps.*** 3rd edition. Springer; 2001.
21. Santamaría R, Therón R: **Treevolution: visual analysis of phylogenetic trees**, *Bioinformatics* 2009; **25:**1970-1971. doi:10.1093/bioinformatics/btp333
22. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, López R, McWilliam H, Remmert M, Söding J, Thompson JD, and Higgins DG: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega**. *Mol Syst Biol* 2011; **7:**539. doi:10.1038/msb.2011.75
23. Cárdenas MI, Vellido A, König C, Alquézar R and Giraldo J: **Exploratory Visualization of Misclassified GPCRs from Their Transformed Unaligned Sequences Using Manifold Learning Techniques.** In *Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering IWBBIO* 2014; **1:** 623-630.
24. Kuang D, Yao Y, Wang M, Pattabiraman N, Kotra LP, Hampson, DR: **Molecular similarities in the ligand binding pockets of an odorant receptor and the metabotropic glutamate receptors**. *J Biol Chem* 2003; **278**(43): 42551-42559. doi:10.1074/jbc.M307120200
25. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M: **The Pfam protein families database**. *Nucleic Acids Res* 2014, Database Issue **42**:D222-D230. doi:10.1093/nar/gkt1223

ENDNOTES

[a] http://www.gpcr.org/7tm/
[b] http://vis.usal.es/treevolution
[c] http://processing.org
[d] http://www.uniprot.org/uniprot/A8DZ72
[e] http://www.ebi.ac.uk/ena
[f] http://www.ncbi.nlm.nih.gov/protein/XP_001515899.2
[g] http://www.ncbi.nlm.nih.gov/protein/XP_002740613

[h] http://www.ncbi.nlm.nih.gov/protein/XP_002123664
[i] http://www.ncbi.nlm.nih.gov/protein/XP_002738008