UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

A computational intelligence analysis of G protein-coupled receptor sequinces for pharmacoproteomic applications

Martha Ivón Cárdenas Domínguez

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament de Ciències de la Computació

A Computational Intelligence Analysis of G Protein-Coupled Receptor Sequences for Pharmacoproteomic Applications

Doctoral Thesis

Author:
Martha Ivón Cárdenas Domínguez

Advisors: **Dr. Alfredo Vellido** and **Dr. Jesús Giraldo**

Artificial Intelligence PhD Program

Barcelona, 2017

To Ivón, Laia, Alan and my loving parents.

Acknowledgements

This dissertation would not have been possible without the guidance and the help of my supervisors Dr. Alfredo Vellido and Dr. Jesús Giraldo. In addition, I would like to thank to the support offered by the projects TIN2012-31377 and SAF2010-19257 of the Spanish Ministry of Economy and Competitivity (MINECO) and research project 110230 from *Fundació La Marató de TV3*.

Abstract

Arguably, drug research has contributed more to the progress of medicine during the past decades than any other scientific factor. One of the main areas of drug research is related to the analysis of proteins. The world of pharmacology is becoming increasingly dependent on the advances in the fields of genomics and proteomics. This dependency brings about the challenge of finding robust methods to analyze the complex data they generate. Such challenge invites us to go one step further than traditional statistics and resort to approaches under the conceptual umbrella of artificial intelligence, including machine learning (ML), statistical pattern recognition and soft computing methods. Sound statistical principles are essential to trust the evidence base built through the use of such approaches. Statistical ML methods are thus at the core of the current thesis.

More than 50 % of drugs currently available target only four key protein families, from which almost a 30 % correspond to the G Protein-Coupled Receptors (GPCR) superfamily. This superfamily regulates the function of most cells in living organisms and is at the centre of the investigations reported in the current thesis. No much is known about the 3D structure of these proteins. Fortunately, plenty of information regarding their amino acid sequences is readily available. The automatic grouping and classification of GPCRs into families and these into subtypes based on sequence analysis may significantly contribute to

ascertain the pharmaceutically relevant properties of this protein superfamily.

There is no biologically-relevant manner of representing the symbolic sequences describing proteins using real-valued vectors. This does not preclude the possibility of analyzing them using principled methods. These may come, amongst others, from the field of statistical ML. Particularly, kernel methods can be used to this purpose. Moreover, the visualization of high-dimensional protein sequence data can be a key exploratory tool for finding meaningful information that might be obscured by their intrinsic complexity.

That is why the objective of the research described in this thesis is twofold: first, the design of adequate visualization-oriented artificial intelligence-based methods for the analysis of GPCR sequential data, and second, the application of the developed methods in relevant pharmacoproteomic problems such as GPCR subtyping and protein alignment-free analysis.

Resumen

Se podría decir que la investigación farmacológica ha desempeñado un papel predominante en el avance de la medicina a lo largo de las últimas décadas. Una de las áreas principales de investigación farmacológica es la relacionada con el estudio de proteínas. La farmacología depende cada vez más de los avances en genómica y proteómica, lo que conlleva el reto de diseñar métodos robustos para el análisis de los datos complejos que generan. Tal reto nos incita a ir más allá de la estadística tradicional para recurrir a enfoques dentro del campo de la inteligencia artificial, incluyendo el aprendizaje automático y el reconocimiento de patrones estadístico, entre otros. El uso de principios sólidos de teoría estadística es esencial para confiar en la base de evidencia obtenida mediante estos enfoques. Los métodos de aprendizaje automático estadístico son uno de los fundamentos de esta tesis.

Más del 50 % de los fármacos en uso hoy en día tienen como "diana" apenas cuatro familias clave de proteínas, de las que un 30 % corresponden a la super-familia de los *G-Protein Coupled Receptors* (GPCR). Los GPCR regulan la funcionalidad de la mayoría de las células y son el objetivo central de la tesis. Se desconoce la estructura 3D de la mayoría de estas proteínas, pero, en cambio, hay mucha información disponible de sus secuencias de amino ácidos. El agrupamiento y clasificación automáticos de los GPCR en familias, y de éstas a su vez

en subtipos, en base a sus secuencias, pueden contribuir de forma significativa a dilucidar aquellas de sus propiedades de interés farmacológico.

No hay forma biológicamente relevante de representar las secuencias simbólicas de las proteínas mediante vectores reales. Esto no impide que se puedan analizar con métodos adecuados. Entre estos se cuentan las técnicas provenientes del aprendizaje automático estadístico y, en particular, los métodos *kernel.* Por otro lado, la visualización de secuencias de proteínas de alta dimensionalidad puede ser una herramienta clave para la exploración y análisis de las mismas.

Es por ello que el objetivo central de la investigación descrita en esta tesis se puede desdoblar en dos grandes líneas: primero, el diseño de métodos centrados en la visualización y basados en la inteligencia artificial para el análisis de los datos secuenciales correspondientes a los GPCRs y, segundo, la aplicación de los métodos desarrollados a problemas de farmacoproteómica tales como la subtipificación de GPCRs y el análisis de proteinas no-alineadas.

Esta tesis se ha desarrollado como parte integral del proyecto de investigación “Adquisición de conocimiento en farmacoproteomica mediante métodos avanzados de inteligencia artificial” (KAPPA AIM) [TIN2012-31377], financiado públicamente a través del MINECO.

Resum

Es podria dir que la recerca en fàrmacs ha tingut un paper predominant en l'avanç de la medicina durant les últimes dècades. Una de les àrees principals de recerca farmacològica és la relacionada amb l'estudi de proteïnes. La farmacologia depèn cada vegada més dels avanços en genòmica i proteòmica, la qual cosa implica el repte de trobar mètodes robusts per a l'anàlisi de les dades complexes que generen. Tal repte ens incita a anar més enllà de l'estadística tradicional per recórrer a enfocaments del camp de la intel·ligència artificial, incloent l'aprenentatge automàtic i el reconeixement de patrons estadístic, entre uns altres. L'ús de principis sòlids de teoria estadística és essencial per confiar a la base d'evidència obtinguda mitjançant aquests enfocaments. Els mètodes d'aprenentatge automátic estadístic seran un dels fonaments d'aquesta tesi.

Més del 50 % dels fàrmacs tenen com a "diana" amb prou feines quatre famílies clau de proteïnes, de les quals un 30 % corresponen a la superfamilia dels G-Protein Coupled Receptors (GPCR). Els GPCR regulen la funcionalitat de la majoria de les cèl·lules i seran l'objectiu central del projecte. Es desconeix l'estructura 3D de la majoria d'aquestes proteïnes, però en canvi hi ha molta informació disponible de les seves seqüències d'amino àcids. L'agrupament i classificació automàtics de les GPCR en famílies i aquestes al seu torn en subtipos sobre la base de les seves seqüències, poden contribuir de forma significativa a

dilucidar aquelles de les seves propietats farmacològicament rellevants.

No hi ha forma biològicament rellevant de representar les seqüències simbòliques de les proteïnes mitjançant vectors reals. Això no impedeix que es puguin analitzar amb mètodes adequats. Entre aquests s'expliquen tècniques provinents de l'aprenentatge automàtic estadístic i, en particular, mètodes kernel. D'altra banda, la visualització de seqüències de proteïnes d'alta dimensionalidad pot ser una eina clau per a l'exploració i anàlisi de les mateixes.

L'objectiu central del projecte será doncs dual: D'una banda, el disseny de mètodes basats en la intel·ligència artificial orientats a la visualització per a l'anàlisi de dades seqüencials corresponents a GPCRs. D'altra banda, i atès que aquesta recerca té la fi última de ser útil en el disseny de fàrmacs i en la comprensió dels processos moleculars involucrats, pretenem aplicar els mètodes desenvolupats a problemes de farmacoproteòmica tals com la subtipificació de GPCRs i l'anàlisi de proteïnes no-alineades.

Acronyms and Abbreviations

AA	Amino Acid
AAC	Amino Acid Composition
ACC	Auto Cross Covariance
AI	Artificial Intelligence
BMU	Best-matching unit
BLOSUM62	Blocks of Amino Acid Substitution Matrix 62
CNS	Central Nervous System
CS	Computer Science
CI	Computational Intelligence
CV	Cross-Validation
DC	Distribution Consistency
DSC	Distance Consistency
DL	Deep Learning
DR	Dimensionality Reduction
EBM	Entropy-Based Measures
EM	Expectation Maximization
EL1	Extracellular loop 1
EL2	Extracellular loop 2
EL3	Extracellular loop 3
ECD	Extracellular Domain
FFT	Fast Fourier Transformation
GTM	Generative Topographic Mapping
GPCR	G Protein-Coupled Receptor

GRAFS	Glutamate Rhodopsin Adhesion Frizzled/Taste2 Secretin
HMM	Hidden Markov Model
IT	Information Technologies
IL1	Intracellular loop 1
IL2	Intracellular loop 2
IL3	Intracellular loop 3
KGTM	Kernel Generative Topographic Mapping
KAPPA AIM	Knowledge Acquisition in Pharmacoproteomics using Advanced Artificial Intelligence Methods
ML	Machine Learning
M-L	Maximum Likelihood
mGluR	Metabotropic Glutamate Receptor
MINECO	Ministerio de Economia, Industria y Competitividad
MSA	Multiple Sequence Alignment
MVD	Multivariate Data
NLDR	Non Linear Dimensionality Reduction
PM	Personalized Medicine
PT	Phylogenetic Tree
PCA	Principal Component Analysis
SOM	Self-Organizing Maps
7TM	Seven Transmembrane
SOCO	Soft Computing
SVM	Support Vector Machine
3D	Three dimensional
TM	Transmembrane
UAB	Universitat Autònoma de Barcelona
UPC	Universitat Politècnica de Catalunya
VFT	Venus Fly Trap

Preface

This Dissertation describes research carried out in the context of the Ph.D. program in *Artificial Intelligence* of Universitat Politècnica de Catalunya (UPC BarcelonaTECH), where the author is part of the Soft Computing (SOCO) research group at the Department of Computer Science (`www.cs.upc.edu`). This research also involved collaboration with the Systems Pharmacology and Bioinformatics research group at Universitat Autònoma de Barcelona (UAB).

The main goals of the Thesis are, on the one hand, the exploration of the potential relationships between receptors of different subfamilies in order to help to identify the multiple roles for individual GPCRs and, on the other hand, GPCR classification analyses based on the physico-chemical properties of the constituent sequence amino acids (AAs) and the contribution of various parameters to the understanding of the mechanisms of specific coupling between GPCRs.

The reported research has been organized in two main parts: **Materials and Methods** (Part I) and **Experiments: Settings, Results and Discussion** (Part II). Its detailed graphical structure is described in figure 1.

Chapter 1 introduces the context of the research, its motivation, and some general concepts that might ease the understanding of the work. This Chapter also presents the main research hypotheses, goals, methodology, the state of the art, the state of research in related areas and a summary of research contributions.

The Materials and Methods part in structured in Chapters 2 and 3:

Chapter 2 describes the analyzed GPCR data and all the data transformations applied in the analyses.

Chapter 3 presents the state of the art of the Computational Intelligence Analysis of GPCRs for pharmacoproteomic applications. In this Chapter, a detailed description of GPCRs and the Machine Learning (ML) techniques proposed for data visualization are provided.

Chapters 4, 5 and 6 describe the Experiments: Settings, Results and Discussion part, where a compilation of the experimental results is provided together with their discussion and corresponding conclusions.

Chapter 4 involves the grouping and visualization of Class C GPCRs family types, resulting in publications number 3, 4, 5, 6, 8 and 9 as listed in Table 1.3. Chapter 5 shows the Grouping and Visualization of mGlu Class C subtypes applied in publications 4 and 7, listed in the same table. Chapter 6 presents an analysis and visualization of error classifications of Class C GPCRs family types, described in publications 4 and 9. All these publications are further commented in section 1.6 from Chapter 1.

Chapter 7 concludes the Dissertation with a commented summary of the results and ideas for further research directions. Finally, the work described throughout the Thesis is matched with the objectives defined in Chapter 1.

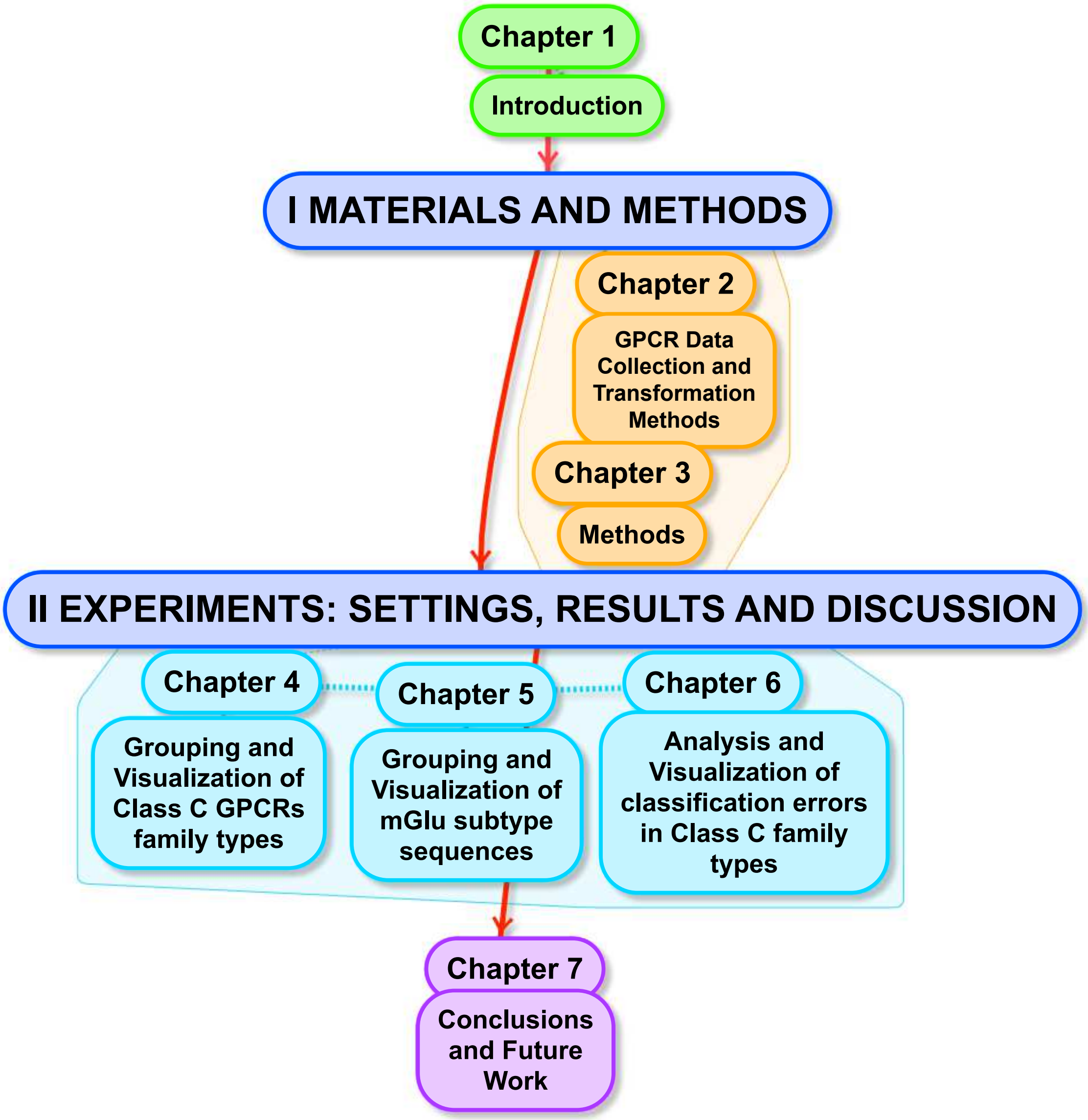

Chapter 1
Introduction
I MATERIALS AND METHODS
Chapter 2
GPCR Data Collection and Transformation Methods
Chapter 3
Methods
II EXPERIMENTS: SETTINGS, RESULTS AND DISCUSSION
Chapter 4
Grouping and Visualization of Class C GPCRs family types
Chapter 5
Grouping and Visualization of mGlu subtype sequences
Chapter 6
Analysis and Visualization of classification errors in Class C family types
Chapter 7
Conclusions and Future Work

Figure 1: Chapters Diagram.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The gap between data generation and data comprehension is widening in all fields of human activity. In medicine and pharmacology, the ambits of this Thesis, the amount of data available for analysis and knowledge extraction is increasing exponentially. The surge in novel techniques for the non-invasive measurement and acquisition of biologically and medically relevant data, in various modalities, is behind this situation. In no other field this is a more pressing challenge than in bioinformatics, where the vast amount of available *omics* information undoubtedly becomes a huge management issue. At the same time, this data *bonanza* should also be understood as a valuable asset for knowledge discovery.

Such gap between data and knowledge must be overcome in order to ensure the advance of biological research and the success of data-based medical decision making support, both key elements of the promised future of personalized medicine.

1.1. Motivation

Arguably, drug research has contributed more to the progress of medicine during the past decades than any other scientific factor. One of the main areas of drug research is related to the analysis of proteins. **Pharmacoproteomics** is the subfield of proteomic research involved in drug discovery and development. It has been argued that it is bound to play an important role in the development of personalized medicine in different ways, including molecular diagnostics [88], [173]. This is because different types of disease, at the level of individual patients and subpopulations, have their own development paths and mechanisms, requiring personalized prevention and treatment. Besides, proteomics can enable the discovery and development of drugs, in a way that is suitable for personalized therapy. Advances in pharmacoproteomics are meant to accelerate the drug development process, and their aims include, amongst others, the verification and identification of drug targets and the elucidation of molecular mechanisms of drug action including efficacy and toxicity [212].

The world of pharmacology is thus becoming increasingly dependent on the advances in the fields of genomics and proteomics, as they should allow us understanding the mechanism of action of a drug. Most modern drug development efforts tend to design compounds that act directly against specific biochemical targets, a task that involves molecular diagnostics, a basis of personalized medicine (PM) [88]. This is an R+D-intensive field with a great potential for knowledge-based economies.

This dependency brings about the challenge of finding robust methods to analyze the complex data they generate. As previously sketched, medicine has become, over the last decade, a data-intensive endeavour. One in which new data-acquisition technologies and a wider variety of investigative goals coalesce to make it one of the most important challenges for multivariate data analysis.

The challenge of managing the complexity of these types of data invites us to go one step further than traditional statistics and resort to approaches under the conceptual umbrella of artificial intelligence, including, amongst others, ML, statistical pattern recognition and soft computing methods, all of which bear the potential to both to scale appropriately to large databases and to deal with non-trivial types of data [122]. In the pursue of this challenge, sound statistical principles are essential to trust the evidence base built through the use of any computational data analysis technique. Statistical ML methods are already establishing themselves in the more general field of bioinformatics [11]. Their use in the area of proteomics have only been reported over the last few years and, as stated in [126], *"proteomics is a much less developed area of research, with data still scarce and fewer computational approaches available for analysis"*, thus, they are at the core of this research work.

At this point, it is worth presenting an historical overview of Spain's pharmaceutical industries. According to Moya-Angeler's report about the Spanish pharmaceutical industry, published in 2008 [137], this industry, 6^{th} in Europe by volume of production (2008 figures) and employing directly over 40,000 people (2006 figures), invested 844 million euros in research in 2006, out of which a 17% was invested in basic research. The sector employs almost 2,500 researchers (as in 2006). These apparently positive figures hide a not so favorable fact: In Europe, the pharmaceutical sector leads the reinvestment of sales into research with a 15.3% (2007 data), surpassing even the IT sector. In Spain, this reinvestment percentage falls to a meager 6.6%. This is mostly due to the comparatively small size and atomization of the Spanish pharmaceutical industry in a global economy context that also leads to the relocation of many R+D centers to emerging economies. Although, according to the study of Industrial Production Index (IPI) of the National Statistics Institute (INE), in 2015 the Spanish pharmaceutical industry grew by 4%, increasing its investment in Spain in 1,000 million euros. In this context, external Spanish research centers, including universities, should play a key complementary role on pharmaceutical investigation at a local level. The lack of enough public-private partnership tradition is in fact another of the barriers limiting this dynamization, as acknowledged in [137]. Breaking this barrier would require the implementation of a national "Strategic Plan for Pharmaceutical R+D+i", similar to others already in place across Europe, such as the seminal "Pharmaceutical Industry Competitiveness Task Force" [158] implemented in the United Kingdom as early as 2000.

More recently, according to the *Cotec 2016 report: Technology and Innovation in Spain* (Fundación Cotec, 2016), the Spanish pharmaceutical industry is at the forefront of the investment in R+D. This sector invested more than 586M € in R+D in 2012 (plus 400M € in external research), and more than 655M € in R+D both in 2013 and 2014, always, more than either the automotive and aeronautics sectors. Only a small percentage of that money (4.1 %) was invested in basic research. In this context, external research centers, including universities, should be able to play a key complementary role, because, as the *Cotec 2016 Report* also reflects, innovation in the pharmaceutic sector strongly depends on investigation carried out in public institutions. Currently, according to **2015 Global Bussiness Reports** publication, Spain is considered as one of the world's most mature pharmaceutical markets, and is scientifically equipped to be an innovative economy. While it has experienced a severe economic crisis-resulting in a decline of the country's pharmaceutical industries, Spain continues to harbor unique potential in terms of its research capabilities and talented pool of human capital, increasing a 30% the relationship between innovating companies, local universities and research centers. [15].

This Thesis is meant to contribute to this basic research with pharmacology as an ultimate target. It focuses on the analysis of a specific type of protein receptors of interest in the field as drug targets. At the cellular level of the central nervous system (CNS), information signal transmission from the extra-cellular to the intra-cellular domain is triggered by receptors. In biochemistry, a receptor can be defined as a protein to which signaling molecules may attach. They are the first step in the process of external signalling, allowing the initiation of intra-cellular signalling cascades after specific ligand binding. Receptors thus play an important role in physiological functions such as cognitive functions: attention, learning, and memory. These functions decline in the course of natural aging and accelerated deficit of cognitive functions is a typical symptom of neurodegenerative diseases such as Parkinson's, Alzheimer's or other pathologies such as anxiety, stress or depression, which are key topics in pharmacoproteomics and the neurosciences.

G Protein-Coupled Receptors (GPCR) are a particular set of membrane-bound receptors and essential components of signal transduction processes in cells. These receptors regulate many cell functions and account for approximately 3% of the human genes. GPCRs act as *antennas* with decoders that allow the cell to understand the signals like a well-tuned radio. If this signal is excessive, inadequate or the receiver is defective, neurological diseases may occur. [167].

The analysis of the gene-family distribution of targets by drug substance reveals that more than 50% of drugs target only four key protein families, from which almost a 30% correspond to the GPCR superfamily. This superfamily regulates the function of most cells in living organisms. Crystal structures are now available for all of the human GPCR classes. They have been reported for more than 60 ligands and 20 receptors, including examples from GPCR classes A, B, C and F. The new structures show previously unobtainable details of interactions between GPCRs and ligands. By June 2014, X-ray structures of 20 different class A, two class B, one class C, and one frizzled GPCR, were available, together with access to their amino acid sequences [31], [189].

These receptors have been the subject of a vast research effort in the pharmaceutical industry due to their ubiquity and involvement in a broad spectrum of physiological functions. Examples of therapeutic indications for drugs acting on GPCRs include: antihistamines, anaesthetics, antidepressants, heart failure, Parkinson's, schizophrenia, migraines and cancer. The design of adequate artificial intelligence-based methods for the analysis of GPCRs will therefore be the focus of this Thesis.

Their automatic grouping and classification into families or classes and these into types and subtypes based on sequence analysis may significantly contribute to ascertain their pharmaceutically relevant properties. There is no biologically-relevant manner of representing the symbolic sequences describing proteins using real-valued vectors. This does not preclude the possibility of analyzing them using appropriate and principled methods. These may come, amongst others, from the field of statistical ML.

Special attention will be paid to the definition of kernels and kernel-based methods, which have become invaluable tools in various fields of data analysis and especially in ML. We aim to obtain specific instances of kernels capable of dealing with sequential structures such as GPCR sequences. The definition of these kernels will open the door to a sensible use of a wealth of well-established and powerful unsupervised learning methods, for the analysis of this kind of data.

Probabilistic modelling and, specifically, probabilistic ML models have only recently begun to be applied to the analysis of GPCRs, although their application is expected to generate new insights in this field. Statistical ML techniques are specially suited to deal with some of the common challenges of molecular modelling in proteins, and should be of special interest at present although some three-dimensional structures of GPCRs have been recently published. The motivation of this research can thus be summarized as the quest for robust probabilistic methods that are capable of grouping and visualizing symbolic protein sequences, on the basis of their structural and functional properties.

1.2. Research Approach

1.2.1. Hypotheses

The research challenges outlined in the previous section motivate us to resort to approaches under the conceptual umbrella of artificial intelligence, including, amongst others, ML, statistical pattern recognition and soft computing methods, for the analysis of GPCR sequences. For the analyses reported in the current study, we resorted to the curated GPCRdb[I] database. In it, GPCRs are divided into five main families, namely Class A, Class B (Secretin), Class C (Glutamate), Adhesion and Frizzled [2]. Recently, there has been a reclassification of receptors into six classes plus the class *Other*: Class A (Rhodopsin), Class B (B1-Secretin and B2-Adhesion), Class C (Glutamate), Class F (Frizzled), Class T (Taste 2) and Class O (Other GPCRs) [138], [83].

In this thesis, we have paid special attention to the GPCR mGluR from the Class C that has generated a wealth of publications over the last few years (a search of the mGluR receptor string in PubMed on December 10^{th}, 2016 produced 164 references only for the year 2016), which is an indication of how attractive they result as pharmacological target for innovative drugs in neurological and psychiatric disorders.

The basic hypotheses of this research can be described hierarchically:

[I] url: http://gpcrdb.org

H0: In the absence of total knowledge about the tertiary and quaternary (3-D models) structure of Class C GPCRs, the overarching working hypothesis is that new knowledge and insights about the structure, subtype characterization and functionality of these receptors can be extracted from the quantitative analysis of both unaligned and aligned primary structure. From this:

H1: Knowledge and insights about the structure, subtype characterization and functionality of Class C GPCRs can be extracted from the analysis of their amino acid sequences directly, that is, analyzing the symbolic sequences, where the symbols belong to amino acid alphabet, and:

H2: Knowledge and insights about Class C GPCRs can be extracted from the analysis of different amino acid sequence transformations, on the basis of either the physico-chemical properties of the amino acids (AAs) or the frequency of repetition of their constituting subsequences.

Again, from these:

H1.1: The direct analysis of Class C GPCR unaligned symbolic sequences will require the development of tailored variations of existing advanced ML methods.

H2.1: The use of probabilistic clustering techniques for the exploratory grouping and visualization, at different characterization levels of the transformed sequences, could reveal interesting insights about receptor subtype structure.

H2.2: The subdivision of the full sequence into biologically meaningful subparts, according to their relative position with respect to the cell membrane, could lead to a better and more applicable receptor subtype structure characterization.

H2.3: The results of Class C GPCR subtype discrimination, as seen from the natural structure of the primary sequence data transformations revealed by unsupervised DR techniques, should differ depending on whether we used the complete primary sequence of these membrane proteins or, instead, we used only the extracellular N-terminus or the 7TM domain.

H2.4: The use of the N-terminus on its own, given the particularities of this domain, should yield comparable results to the complete sequence in terms of subtype discrimination.

and finally,

H3: Protein similarity measures and receptor subtype structure characterizations can be used for transferring the knowledge of structure-activity relationships with small molecules from one Class C GPCR to others.

1.2.2. Objectives

Overall, thus, the central objective of this research work is dual: on one side, the definition, design and implementation of visualization-oriented artificial intelligence-based methods, with sound foundations on statistical theory, for the exploration and analysis of GPCR sequential data. On the other side, and given that this research has the ultimate goal of being useful in helping drug design and in achieving a better understanding of the molecular processes involved in receptor signalling both in normal and pathological conditions, we aim to apply the developed methods for investigating relevant pharmacoproteomic problems such as GPCR subtyping, the exploration of the phenomena of receptor heteromerization and deorphanization, and protein alignment-free analysis.

Based on the hypotheses listed above, two main objectives are outlined:

1. Grouping and visualization of GPCRs.

   The goal of GPCR grouping, where the availability of accurate knowledge of its crystal 3D is limited and, therefore, proteins are specified by their aligned amino acid primary sequences, aims to find biologically meaningful partitions. This might help the analyst to make inferences about key protein regions and residues both in the obtained groups and for the whole family.

   1.1. Grouping and visualization of GPCRs using kernel manifold learning.

   In order to group GPCR sequences, we need a measure of similarity between them. ML techniques can help us in this task. Unsupervised data analysis using clustering algorithms provides a useful tool to explore data structures. Over the last few years, several kernel methods for visualization and clustering of non-standard multivariate data have been proposed. In this Thesis we aim to define appropriate kernels for sequence similarity analysis and to embed them in statistical ML methods of the manifold learning family capable of simultaneous clustering and visualization.

   1.2. Matching of GPCR subtypes and phylogenetic trees.

   A phylogenetic tree is a dendrogram-like graphical representation of the evolutionary relationship between groups of proteins which may share a set of homologous characters. This is a common tool for protein subtyping on the basis of sequences. We have aimed to match the subtyping cluster structure obtained with the methods described in the previous point with that yielded by standard phylogenetic trees.

2. Further pharmacoproteomic challenges.

   Kernel manifold learning methods is used to help in the exploration of receptors with very heterogeneous grouping structure. This heterogeneity might be an indication of their susceptibility towards heterodimerization, which could be useful in the quest of more potent and safer drugs.

2.1. Family subtype exploratory sequence visualization from their constituent sub-parts.

Analysis of the different roles of extra-cellular / transmembrane parts of the receptor on subtype characterization at different levels of detail, helping to find GPCRs susceptible of heterodimerization. We investigate whether the separated analysis of two of the three differentiated receptor domains yields any advantage in terms of reducing the level of overlapping between the apparently more difficult to discriminate subtypes.

2.2. Visual discrimination assessment measures.

The results of unsupervised techniques applied to the visualization and clustering of available Class C GPCR data must be assessed for quality. In order to complement the qualitative exploratory visualization of the Class C GPCR sequences, we describe here several measures for the quantitative assessment of subfamily overlapping that are suitable for discrete clustering visualizations such as those provided by the unsupervised models applied in the thesis.

1.2.3. Limitations

Given that our analyses are based on receptor primary sequences, a first problem for visualization obviously arises: the transformation of varying-length sequential symbolic data into formats that are suitable for multivariate data analysis. Roughly speaking, those transformations might use the complete sequences in unaligned form or might instead apply methods of multiple sequence alignment (MSA). Both are used in our experiments. Although, these transformations lead to a second problem: that of the high dimensionality of the transformed data, making direct data visualization impossible. In this scenario, dimensionality reduction (DR) methods are necessary [117]; out of the many DR families of techniques available to the analyst, our work focuses here on manifold learning methods.

1.3. Methodology

The definition of a methodology allows the explicit enumeration of the steps that need to be followed by the researcher to address a particular problem. In this case, with the hypotheses defined above as a reference and the objectives set as goals, I now explain the methodology of the main tasks:

- **Grouping and visualization of aligned GPCRs using manifold learning**.

  As already mentioned, in absence of total knowledge about their 3-D physical structure, the proteins are specified by their amino acid sequences. In order to get a better understanding of the functional role of the members of a protein family in biochemical processes, it is important to know the internal organization of the family and the detection of key regions where interactions with other molecules may take place or which are essential to inform the 3D structure of the protein.

  This internal organization can be explored using grouping or clustering procedures. In order to group GPCR sequences, we need a measure of similarity between them. In this thesis we will depart from some existing kernel models to a GPCR-specific kernel, as part of a kernel-based statistical ML model of the manifold learning family, namely the Kernel

Generative Topographic Mapping (KGTM). This model describes multivariate data in terms of low dimensional representations, so as to achieve the visualization of high dimensional data that would otherwise be difficult to visualize. The visualization of the high-dimensional GPCR sequences would considerably help understanding their global grouping structure.

We use the probabilistic properties of KGTM to explore GPCRs subclasses in more detail. For that we resort to the explicit calculation of the probability of each of the available sequences belonging to each of the model groupings. This provides us with a map of probability that can qualify the differences between sequences of either clear or dubious subclass ascription.

- **Matching the hierarchy of GPCR subtypes and phylogenetic trees**

  The evolutionary relationship between groups within each of the families in the GPCR super-family remains unknown at large. They may have diverged from a common ancestor, or perhaps be the result of convergent evolution, in which functional constraints push unrelated proteins from different organisms towards the same design. The sequences of different GPCR families are highly diverged from each other, except that they share one common structural feature, namely, they all have seven hydrophobic transmembrane regions.

  Generally speaking, a phylogenetic tree is a dendrogram-like graphical representation of the evolutionary relationship between taxonomic groups. It can also be seen as a specific type of cladogram where the branch lengths are proportional to the predicted or hypothetical evolutionary time between groups. They are not meant to be understood as completely true and accurate descriptions of the evolutionary paths they represent, because in any of them there are a number of possible evolutionary pathways that could produce the pattern of relatedness they represent. In the case of GPCR sequences, they only illustrate the probability that two sequences are more closely related to each other than to a third one.

There are standard phylogenetic tree visualization tools, such as, for instance Jalview 2.6.1, which uses the standard Blocks of Amino Acid Substitution Matrix 62 (BLOSUM62) [74] as a basis.

- **Further pharmacoproteomic challenges: finding GPCRs susceptible of heterodimerization**

  Kernel manifold learning methods could be used to help in the exploration of receptors with very heterogeneous grouping structure. This heterogeneity might be an indication of their susceptibility towards heterodimerization, which could be useful in the quest of more potent and safer drugs. No that much is yet known about the ability of the receptors to interact to form new functional structures. The concept of GPCR heterodimerization, or the physical association of two different types of GPCRs, presents an unexpected mechanism for GPCR regulation and function, and provides a novel target for pharmaceuticals [71].

  Specifically, the heterodimerization of GPCRs is a function-modulating mechanism. We hypothesize that the assignment of GPCR sequences to class-overlapping spaces by either unsupervised or supervised methods can be an indication of their propensity to heterodimerize.

1.4. State of the art: GPCRs

The convergence between proteomics and AI is providing new tools and methods for the discovery and knowledge extraction from the complex data generated in the biology field. Here, we first provide the biological context for the current Thesis, introducing the bioinformatics domain of our work.

The section starts with a brief introduction of the GPCR superfamily, commonly divided into five main families, namely Class A, Class B (Secretin), Class C (Glutamate), Adhesion and Frizzled. Then, the emphasis is placed on our specific area of study, which is the class C of GPCRs.

The idea of receptors has fascinated scientists for more than a century and today the G-protein coupled receptors (GPCRs), also known as seven transmembrane receptors, represent by far the largest, most versatile and most ubiquitous of the several families of plasma membrane receptors. In fact, the ability of cells to communicate with each other using chemical messengers in the form of hormones and neurotransmitters, is in essence, an information encoded using GPCRs located in the plasma membrane.

Despite the very central role that the study of receptors plays in biomedical research today, it is only in the last thirty years that there has been any general acceptance they even exist. Therefore, Dr. Lefkowitz and Dr. Kobilka, awarded with the 2012 Nobel Prize in Chemistry, and pioneers of the biochemical techniques on GPCRs in the early 70s, allowed researchers to study the regulation of the receptors by numerous factors, to discover previously unsuspected receptor subtypes, and to develop theories concerning the mechanisms of receptor action [118], [101]. The following sections describe the main characteristics of the GPCRs and its classification.

1.4.1. GPCRs: structure, function and classification

In biochemistry, a **receptor** is a protein molecule that receives chemical signals from outside a cell, located either on a cell's surface, that binds to a specific ligand (typically an ion or a molecule), initiating signal transduction and a change in cellular activity. Receptors play an important role in physiological functions. This Thesis focuses on **G-Protein Coupled Receptors** (GPCRs), which are membrane receptors that modulate biochemical functions by coupling to and activating G proteins. The name is derived from their association with heterotrimeric G proteins, which have GTPase activity and act as intermediary components, activating or inhibiting several intracellular effectors.

These receptors are very attractive drug targets in the quest for new medicines. They are the largest, most important and best-validated class of pharmaceutical target proteins and here we focus in the class C GPCR family, which is involved in several major CNS disorders [160]. The GPCRs usually share a 7TM helix topology with an extracellular N-terminus and an intracellular C-terminus.

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 [205], [42]. 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 extracellular signal is invariably transduced to a cytosolic heterotrimeric G protein complex. GPCRs can be divided into families with a striking lack of common sequence motifs [23],[56]. This is reflected by the vast number of extracellular ligands that activate the receptors, which range from neurotransmitters, hormones and peptides to external stimuli such as light, taste and odors [203].

The human GPCRs can be classified into six classes, and as many unique (other) receptors (Table 1). Two overlapping classification systems have denoted the classes A-F (Kolakowski, 1994) or by their GRAFS members (**G**lutamate **R**hodopsin **A**dhesion **F**rizzled/Taste2 **S**ecretin), based on sequence homology and phylogenetic analysis (Fredriksson et al., 2003) respectively [84] (See table 1.1). The taste type 2 receptors were recently placed as a separate sixth

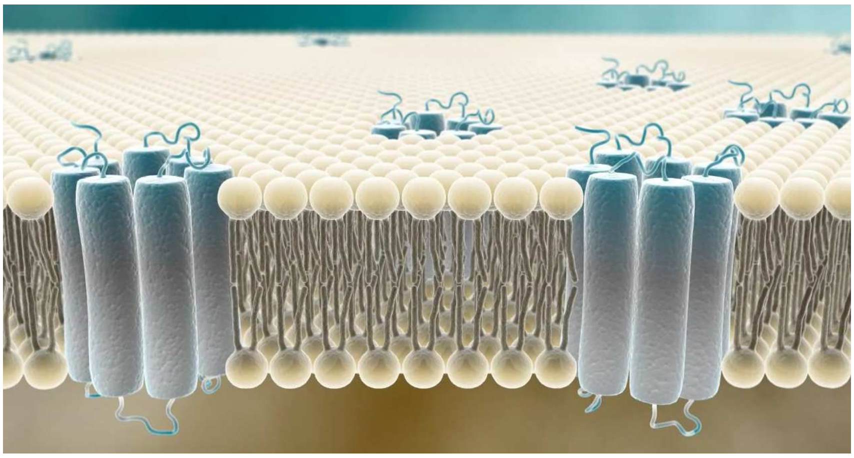


Figure 1.1: GPCR illustrative representation.

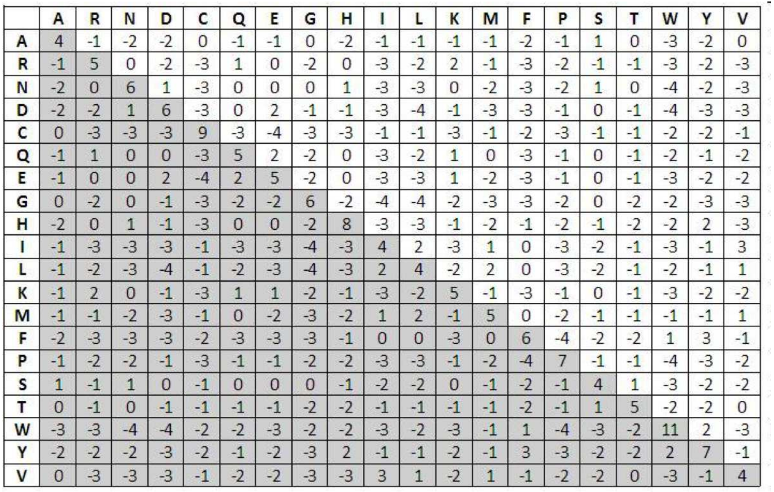
class having evolved from class A (Nordstrom et al., 2011). The classes are further grouped into receptor families by pharmacological classification of their endogenous ligands that span ions, neurotransmitters, lipids, carbohydrates, nucleotides, amino acids (AAs), peptides and proteins (Southan et al., 2016). The pharmacological receptor families mirror the evolutionary subfamilies, with a few exceptions. Recently, there has been a reclassification (not yet standard) of receptors according to the GRAFS (**G**lutamate **R**hodopsin **A**dhesion **F**rizzled/Taste2 **S**ecretin) system which, as the acronym indicates, includes the following groups: glutamate, rhodopsin, adhesion, frizzled/taste2, and secretin [84]. This research uses the GPCRdb classification, which is based on the GRAFS classification system (See table 1.1).

The assumption that similar molecules bind to similar receptors [100] and that small molecules bind within the upper part of the transmembrane helices, gives rise to the application of pattern recognition analysis on multiple sequence alignments of those helices or parts thereof to identify ligand binding residues [172].

The problem of the pairwise sequence alignment is that gaps have been inserted in both the top and the bottom sequence in order to slide the sequences along so that the regions of similarity between them become apparent. There are two possibilities for the alignment: the mutation from a letter to another or vice-verse and the deletion or insertion. In that sense, we are looking for the alignment that we think is most likely to have occurred during evolution.

With the aim of obtaining the best alignment, a scoring system is required. The sequences to be aligned can be represented as letters from a symbolic alphabet of allowed characters. In our case, for proteins, it is the 20-letter alphabet corresponding to the same number of AAs. A useful scoring system for proteins needs to reflect the fact that some AAs are similar to one another whereas others are different.

With that purpose, a high positive score will be assigned to two identical AAs, a slightly positive score to two similar AAs *(e.g. D and E)* and a slightly negative score to a pair of very different AAs [73]. In the scoring matrix process, the optimal alignment produced by the algorithm will depend on the scoring system. Figure 1.2 gives an example of the BLOSUM62 scoring system.



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 1.2: An example of commonly used scoring system: BLOSUM 62.

Table 1.1: G-protein coupled receptor families according to GPCRDdb, based on the GRAFS classification system.

Superfamily	Description
Family A	Receptors related to Rhodopsin-like Receptors
Family B	Receptors related to the Calcitonin and PTH/PTHrP Receptors
Family C	Receptors related to the Metabotropic Glutamate Receptors
Family D	Receptors related to the pheromone Receptors
Family E	Receptors related to the cAMP Receptors

The second part of the scoring system is the penalty for gaps. The likelihood of a single event that creates a gap of residues depends on the underlying mutational process involved according to the divergence of the corresponding amino acid sequences. It also depends on the selection process. Many insertions or deletions may lead to non-functional sequences which may be eliminated by selection. Moreover, there are certain regions of sequences where it is easier to find gaps. For example, unstructured loop regions in proteins tend to be much more variable that well-defined structured regions, such as helices in membrane proteins. [76]

It worth to mention that along the evolutionary history of GPCRs, different classification have been given to the GPCR superfamily, depending of specific criteria. For example, receptors have been grouped by how their ligand binds, or by comparing physiological and structural aspects given their considerable complexity. Fredriksson and colleagues [56], did a great phylogenetic study of the entire superfamily, while Takeda and colleagues [187], determined that over 800 GPCR genes in the human genome fall into three major families *(termed rhodopsin, secretin, and metabotropic glutamate receptor-like)*, which are populated by more than 50 receptor types and 350 subtypes.

In order to manage the heterogeneity of proteomic databases, it is necessary to standardize them for quality assessment, storage and also to ensure accuracy and reproducibility of the data, that is, data curation processes must be implemented.[45]

Despite the diversity of the superfamily, certain commonalties remain within all GPCRs. All proteins within the GPCR superfamily contain seven highly conserved transmembrane segments which are highly hydrophobic. Sequences can therefore be divided into the regions (see figure 1.3 and figure 1.4 for a top view of the previous one) described in Table 1.2.

Table 1.2: Representation of the regions of a GPCR sequence: three extracellular loops (EL1,EL2,EL3), three intracellular loops (IL1,IL2,IL3), seven transmembrane regions (TM1 → TM7) and the protein termini (N-terminus (NT) and C-terminus (CT)). The arrangement of the connected GPCR regions is conserved across the three domains starting with the NT and ending with the CT.

GPCR Regions Arrangement
NT-TM1-IL1-TM2-EL1-TM3-IL2-TM4-EL2-TM5-IL3-TM6-EL3-TM7-**CT**

Moreover, the identification of the transmembrane regions (TM1, TM2, TM3, TM4, TM5, TM6, TM7) also informs the remaining structure of the GPCR.

The considerable complexity accumulated along the evolutionary history of GPCRs implies that members of different families share almost no recognizable sequence similarity, despite being linked by a similar 7TM architecture. The receptors are activated by different ligands (ions and small and large molecules such as proteins), some of binding sites are located in the external loop regions and some in the internal 7TM. The activated receptors then interact with different G proteins or other intracellular molecules to effect their diverse biological responses.

The difficulties of understanding and analyzing computationally the richness of their evolutionary relationships and the complexity of their interactions with other molecules are well reported. Many tools have been used to reduce the

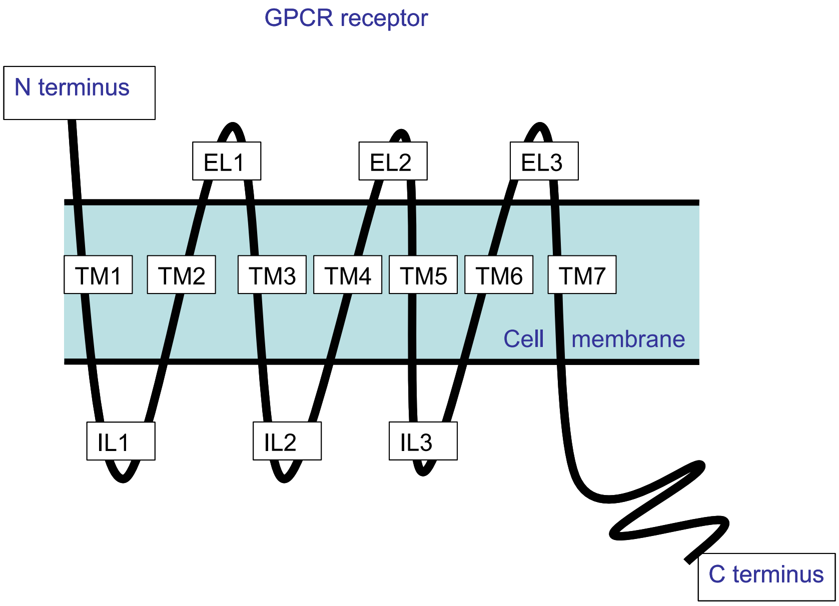


Figure 1.3: GPCR common structure: The transmembrane segments form seven α-helices in a flattened two-layer cell membrane. The transmembrane regions are: TM1, TM2, TM3, TM4, TM5, TM6 and TM7.

complex problem of sequence relationship analysis. Unfortunately, identifying relationships between sequences is clearly not the same as identifying their functions. In that sense and as an example, computational approaches can be useful for receptor deorphanization, that is, for the characterization *ex novo* of those receptors whose endogenous ligand is unknown [66],[8].

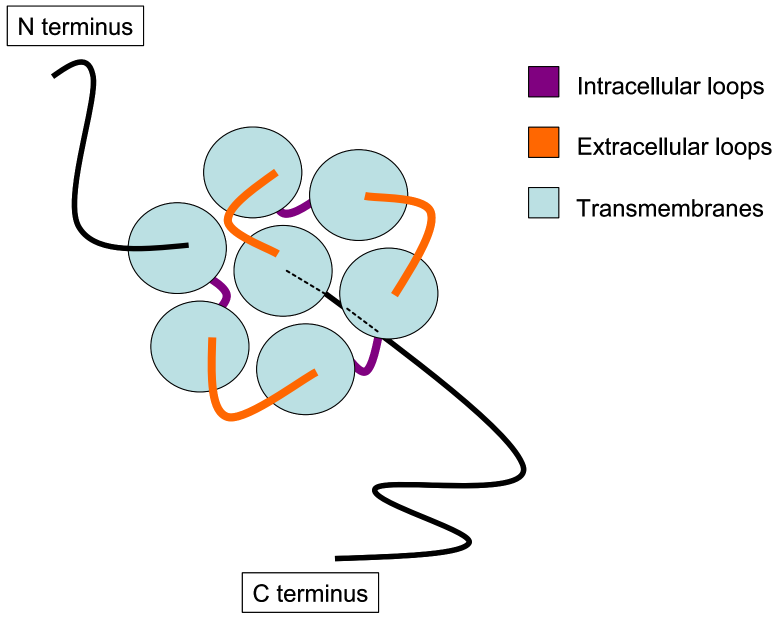


Figure 1.4: Schematic top-view of a GPCR where is visualized the spatial disposition in spiral of the seven transmembranes, the three intracellular loops and the three extracellular loops along the cell membrane.

1.4.2. GPCR sequences homology and heterodimerization

The concept of homology leads to a specific classification of the evolutionary relationships between members of protein families. Sequences or structures are homologous if they are related by evolutionary divergence from a common ancestor [76]. It means that homology cannot be directly observed, but can be inferred from calculated levels of sequence or structural similarity. In effect, reliable threshold values of similarity are dependent on the mathematical methods used for analysis. Homologous proteins can be recognized by sequence comparison because strong selective constraints prevent amino acid substitutions in particular positions from being accepted. The methods developed for comparing protein sequences seek to infer homology on the basis of the correct alignment of the residues between proteins attempting to determine all the equivalent residue positions.

The concept of GPCR heterodimerization was initially proposed in the early 1980's [1]. Heterodimerization seems to be selective, so that GPCRs will interact with one type of receptors, but not with others. GPCRs have traditionally been thought to act as monomers, but this idea has been challenged over the past few years by accumulating pharmacological and biochemical data [109]. However, many investigations has evidenced that GPCRs may physically interact with each other and that oligomeric forms of the same receptor (homodimers) or different receptors (heterodimers) may be functionally active [24], [132],[3], [124], [168], [97].

In that sense and focusing in class C of GPCRs, many researchers provided general results about the heterodimerization in some receptors. Gama and Colleagues [60], investigated how the Calcium Sensing receptors (CaR) and the group I methabotropic glutamate receptors (mGluR1 and mGluR5) can form heterodimers and also reported their functional positive interaction.

GPCR heterodimerization can in some cases alter the pharmacological properties of the associated receptors, such that novel pharmacological entities are created. Since GPCRs are important drug targets in the treatment of many different diseases, understanding the specificity and physiological significance of GPCR heterodimerization may lead to insights that will fundamentally impact the development of future therapeutics, having a great importance since GPCRs are the molecular targets for numerous therapeutic drugs [161].

1.4.3. Functional GPCR relationships

One of the recent findings in GPCR research is the clarification of the mechanisms of GPCR oligomerization at a molecular level. Many studies have suggested the importance of transmembrane α -helices for GPCR oligomerization. The common elements of their structural and functional features are responsible for the presence of detectable patterns of motifs and correlated mutations that may be revealed from the alignment of the sequences of these complex biological systems. The decoding of these patterns in terms of structural and functional determinants can provide indications about the most likely interfaces of dimerization/oligomerization of GPCRs [51].

The most widely used strategy to link sequence or structure to function, namely homology-based function prediction, relies on the fundamental assumption that sequence or structural similarity implies functional similarity [6]. To address this, the application of sequence similarity networks for visualizing functional trends across protein superfamilies from the context of sequence similarity, can be used. Figure 1.5 shows an example of a GPCR network visualization:

Cytoscape software [36] can be used for bringing data together under a graphical network paradigm. Cytoscape provides an *in silico* approach to examine and display protein interaction networks based on available protein-protein interactions characterized from previous biological studies [81], [178]. These networks contain the proteins of a family with distances calculated from the family alignments. For all proteins, the protein family information, species names and the amino acid types for all the residues annotated with a general residue number are available as attributes. This allows for complex analyses, such as coloring proteins by AAs at a certain residue position to compare species or sub-type specific differences.

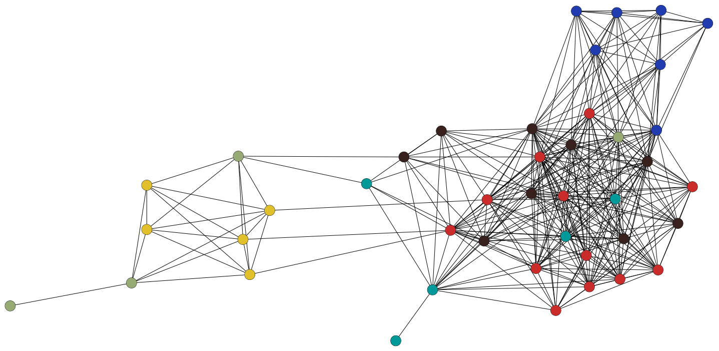


Figure 1.5: Sequence similarity network including the including Amine-binding and Class A GPCRs. With Cytoscape software, nodes and edges can have attributes associated with them and subnetworks can be extracted and scored.

1.4.4. Class C GPCRs: the focus of our research

Class C of the GPCR superfamily has become an increasingly important target for new therapies, particularly in areas such as pain, anxiety, neurodegenerative disorders and as antispasmodics, but also potentially for the treatment of hyperthyroidism and osteoporosis [68].

This class represents a distinct group of the GPCR superfamily, having a specific extracellular domain known as the Venus flytrap (VFT), which is responsible for ligand recognition and binding (see figure 1.6). Its conserved 7TM domain is characterized by intracellular loops and a C-terminal domain. Interestingly, although the ligand binding site is located in the N-terminal domain, the receptors can be modulated by allosteric modulators (positive allosteric modulators (PAMs) and negative allosteric modulators (NAMs)) binding to the 7TM domain. [65]

The research in this thesis will pay special attention to metabotropic gluta-

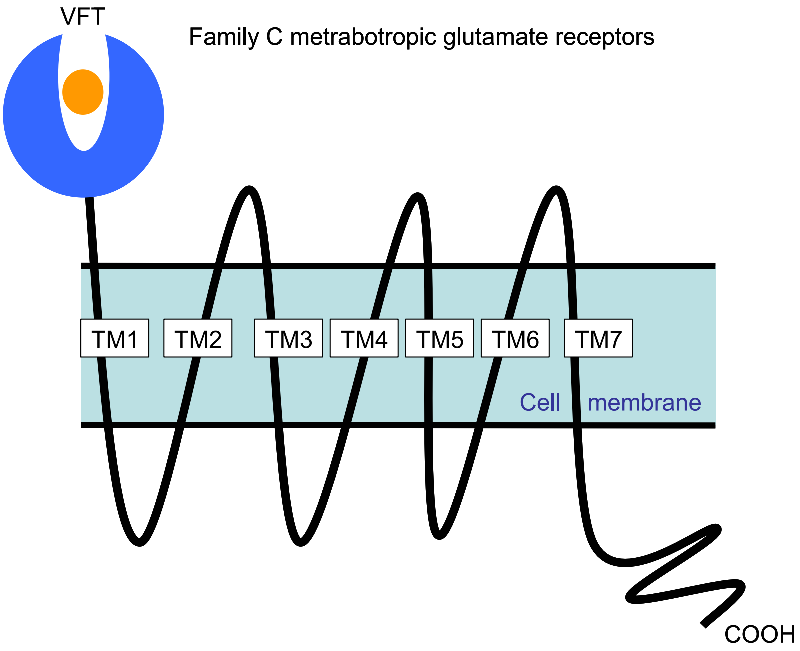


Figure 1.6: GPCR class C: the 7 loops representing the 7 transmembrane regions; VFT is the Venus Fly Trap and COOH is the C terminus.

mate (mGlu) receptors. The mGlu receptors, which belong to the first group of GPCR class C, are activated by glutamate, the major excitatory neurotransmitter in the central nervous system, and play important roles in regulating cell excitability and synaptic transmission. The mGlu receptors are widely distributed throughout the CNS, and a whole range of neurological and psychiatric disorders might be treated using drugs that act directly on these receptors.

The wide diversity and heterogeneous distribution of mGlu subtypes provides an opportunity for selectively targeting individual mGlu subtypes involved in only one or a limited number of CNS functions for the development of novel treatment strategies for psychiatric and neurological disorders.

Class C GPCRs have a rich and deep taxonomy of subtypes. As an example, one of class C subtypes (according to GPCRdb), namely Metabotropic Glutamate Receptors, is further subdivided into eight types (type 1 to type 8), where type 3, for instance, is in turn subdivided into 6 subtypes (1 to 6). As a result of this rich characterization, the automatic discriminatory classification of their subtypes becomes a non-trivial problem that often requires the use of multivariate data analysis tools.

Note though that this rich characterization is not limited to class C. As Gao and colleagues [63] demonstrated, GPCR subtyping can be performed at up to seven increasingly specific levels of detail, from the general level of GPCRs *vs.* non-GPCR proteins all the way down to the most specific characterized subgroups. These results built on previous GPCR hierarchical classification attempts in [39, 62, 155]. All these studies provide general results at each level of GPCR representation, but no specific results for class C subtyping, so that they cannot be directly compared to those obtained in the current study.

The problem of primary sequence-based GPCR classification has been investigated from the last decade of the 20^{th} century [104], [39]. Computational intelligence and machine learning approaches have become popular in this domain and Support Vector Machines (SVM), in particular, are the method of choice in many studies [39, 63, 186, 166, 107, 82], including some in which the type of alignment-free data transformation is not too different from those used in this study [123, 90]. They are also the analytical building block of GPCR classification software tools such as PRED-GPCR [70], GPCRPred [17] and of software tools for the related problem of homology detection from sequences such as, amongst others, SVM-I-sites [78], SVM-BALSA [201], SVM-n-peptide [145] and SVM-HUSTLE [177].

Few studies have been devoted to unsupervised approaches to the analysis of GPCRs from the point of view of subtype discrimination. Some interesting examples can be found in [114, 148]. In the work by Lapinsh and colleagues [114], Principal Component Analysis was applied by calculating principal components of physicochemical properties of the amino acids in the sequences. In [148], Self-Organizing Maps (SOM: a type of unsupervised artificial neural networks) were used to cluster and visualize unaligned sequences, aiming to discriminate between subtypes. This last approach is specially relevant to our research, given that the methods described in Chapter **Data Visualization** are functionally-similar probabilistic alternatives to SOM. Both of these studies analyzed sequences from class A of GPCRs.

Comparatively, little research has specifically been devoted to the sequence-based classification of class C GPCR subtypes. An advanced variant of Hidden Markov Models was used in [185] to discriminate subtypes of different GPCR families including some of those in class C, using data also analyzed in [90]. Despite the data in both studies were extracted from the GPCRdb repository (described in section 2.1), unfortunately, results are not directly comparable for two reasons: the data in [185, 90] were acquired from the 2000 database version (more than a decade older than the one we analyzed) and the evaluation metrics are completely different. A fast Fourier transformation of classes B, C, D and F sequences was used in [96] to classify GPCR subtypes using Nearest Neighbor classifiers. Data were acquired from a 2006 version of the GPCRdb. Only 403 sequences from all classes were available. Semi-supervised classification of unaligned sequences was performed in [34] with the goal of sequence de-orphanization.

1.5. State of research in related areas

Focusing this description in Spain, there exist few groups in the area of AI with bioinformatics as the main acknowledged application area (an area in which pharmacoproteomics would be inscribed). Some of the main ones include Profs. Bielza's and Larrañaga's Computational Intelligence Group (CIG) at Universidad Politécnica de Madrid, whose research includes proteomics [64], [129]; Dr. J.S. Aguilar-Ruiz's Bioinformatics Research Group at Universidad Pablo de Olavide, Sevilla, working mostly in gene expression data analysis and protein contact prediction [128]; Dr. Lozano's Intelligent Systems Group in collaboration with the Biomics Research Group at the Universidad del País Vasco [14]; and the M4MLab, part of Prof. F. Herrera's Soft Computing and Intelligent Information Systems (SCI2S) research group at the Universidad de Granada [32].

To the best of our knowledge, no Spanish research group in the field of AI has pharmacoproteomics as a main application area, with exceptions such as Dr. García Rodríguez's work [26] at the I2RC research group of *Universidad de Alicante*. We should add to this the sporadic forays of groups associated to the *Instituto Nacional de Bioinformática* (`www.inab.org`)

Now broadening the scope, diverse approaches for classifying GPCRs that resort to ML techniques have been proposed. One from which we could draw inspiration for some of our developments [209] has recently been proposed by researchers from the *Division of Medical Chemistry* at Leiden University and the Department of Computer Science at Vrije Universiteit, in The Netherlands. Receptor classification has been studied with SVMs by Strope and Moriyama at the University of Nebraska-Lincoln, USA [25].

Other applications of ML in the field, mostly related to the analysis of

GPCRs, include prediction of receptor binding sites and virtual screening at the *Structural Chemogenomics Group*, Université de Strasbourg, France [202]; multiplicity and selectivity of GPCR-G protein interaction at the *Computational Biology Research Center*, National Institute of Advanced Industrial Science and Technology in Tokyo, Japan [207], prediction of receptor residues involved in protein interactions, at the University of Toronto, in Canada (Hui et al 2013) and at the *Korea Research Institute of Bioscience and Biotechnology* [46]; or analysis of multiplicity and selectivity of GPCR-G protein interaction at *AIST* in Tokyo, Japan.

Spanish research groups working on GPCRs from bioinformatics (even if not CI-related) perspective, include two groups working at UAB: the *Laboratory of Molecular Neuropharmacology and Bioinformatics* and the *Laboratori de Medicina Computational*. Others include the *Grup de Recerca en Bioinformàtica i Estadística Mèdica*, at *Universitat de Vic*; the Computer-Assisted Drug Design Lab [25] at *Universitat Pompeu Fabra* (UPF), led by Dr. Pastor, and more indirectly, Dr. Ismel Brito's work at IIIA-CSIC, in collaboration with Dr. Borroto-Escuela at the *Karolinska Institutet* in Sweden. Internationally, CI and ML have become a standard for data analysis in bioinformatics [120], but their application to GPCR analysis is still limited. A brief selection of recent approaches are listed next: Analysis of specificity determining residues from MSA [48]; receptor classification using SVM has, for instance, been investigated in [70]; [186]; [177] and more recently in [89]; [135], for instance. A review on classification methods for the analysis of GPCRs can be found in [206].

1.6. Contributions

Over the completion of this Thesis, some of the reported research was published. Part of it became the M.Sc Thesis that preceded the current document: **Kernel-based manifold visualization of GPCR sequences**. The author contributed in the publications listed below:

- **VCORG2011**
  Year: 2011
  Title: **A probabilistic approach to the visual exploration of G Protein-Coupled Receptor sequences**.
  Authors: Alfredo Vellido, Martha Ivón Cárdenas, Ivan Olier, Xavier Rovira and Jesús Giraldo.
  Type publication: Congress - *In Proceedings of the 19th European Symposium on Artificial Neural Networks (ESANN), pp.233-238.*

- **CVORG2012**
  Year: 2012
  Title: **Complementing Kernel-Based Visualization of Protein Sequences with Their Phylogenetic Tree**.
  Authors: Martha Ivón Cárdenas, Alfredo Vellido, Ivan Olier, Xavier Rovira and Jesús Giraldo.
  Type publication: Congress - *LNCS/LNBI 7548, pp.136-149.*

- **CVORG2012a**
  Year: 2012
  Title: **Kernel Generative Topographic Mapping of Protein Sequences** Authors: Martha Ivón Cárdenas, Alfredo Vellido, Ivan Olier,

Xavier Rovira and Jesús Giraldo.
Type publication: Book Chapter - *Medical Applications of Intelligent Data Analysis: Research Advancements.* In: R. Magdalena-Benedito, E. Soria, J. Guerrero Martínez, J. Gómez-Sanchis and A.J. Serrano-López (eds.) IGI Global, pp.194-207, doi: 10.4018/978-1-4666-1803-9.

- **CVKAG2014**
  Year: 2014
  Title: **Exploratory visualization of misclassified GPCRs from their transformed unaligned sequences using manifold learning techniques**
  Authors: Martha Ivón Cárdenas, Alfredo Vellido, Caroline König, Rene Alquézar and Jesús Giraldo.
  Type publication: Congress - *Proceedings of the 2nd International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2014) pp.623-630.*

- **CVG2014**
  Year: 2014
  Title: **Visual interpretation of class C GPCR subtype overlapping from the nonlinear mapping of transformed primary sequences.**
  Authors: Martha Ivón Cárdenas, Alfredo Vellido and Jesús Giraldo.
  Type publication: Congress - *In Proceedings. of the 2nd International Conference on Biomedical and Health Informatics (IEEE BHI) pp.764-767.*

- **CVG2014a**
  Year: 2014

Title: **Exploratory visualization of Metabotropic Glutamate Receptor subgroups through manifold learning.**

Authors: Martha Ivón Cárdenas, Alfredo Vellido and Jesús Giraldo.

Type publication: Congress - *17th International Conference of the Catalan Association of Artificial Intelligence (CCIA) In L. Museros et al. (Eds.) Artificial Intelligence Research and Development, IOS Press, pp.269-272.*

- **CVKAG2015**

  Year: 2015

  Title: **Visual Characterization of Misclassified Class C GPCRs through Manifold-based Machine Learning Methods.**

  Authors: Martha Ivón Cárdenas, Alfredo Vellido, Caroline König, Rene Alquézar and Jesús Giraldo.

  Type publication: Journal - *Genomics and Computational Biology, 1(1) e19 .*

- **KCGAV2015**

  Year: 2015

  Title: **Label noise in subtype discrimination of class C G-protein coupled receptors: A systematic approach to the analysis of classification errors.**

  Authors: Caroline König, Martha Ivón Cárdenas, Jesús Giraldo, Rene Alquézar and Alfredo Vellido.

  Type publication: Journal - *BMC Bioinformatics, 16(1):314.*

- **CVG2016**

  Year: 2016

  Title: **Visual exploratory assessment of class C GPCR extracellular domains discrimination capabilities**.

  Authors: Martha Ivón Cárdenas, Alfredo Vellido, Jesús Giraldo.

  Type publication: Congress - *The 10th International Conference on Prac-*

tical Applications of Computational Biology & Bioinformatics (PACBB) Advances in Intelligent Systems and Computing 477, pp.31-40.

Table 1.3: Summary of Contributions according to Application Tasks (A), Processed Data (D), Data Transformation (T) and Applied Methods (M).

Num	Publication	A	D	T	M
1	VCORG2011	C232 mGluR	Full-Seq	MSA	KGTM
2	CVORG2012	C232	Full-Seq	MSA	KGTM PT
3	CVORG2012a	C232 mGluR	Full-Seq	MSA	KGTM
4	CVKAG2014	C1,510 Error	Full-Seq	MSA AAC	GTM PT
5	CVG2014	C1,510	Full-Seq	AAC MSA	GTM KGTM,EBM
6	CVG2014a	C1,510 mGluR	Full-Seq	AAC,DI MSA	GTM, KGTM,EBM
7	CVKAG2015	C1,510 Error	Full-Seq	AAC MSA	GTM PT
8	KCVAG2015	C1,510 Error	Full-Seq	AAC MSA	GTM KGTM,PT
9	CVG2016	C1,510	Full-Seq EC-Dom	AAC MSA	GTM KGTM,EBM

The description of A, D, T and M in the previous table is as follows:

1. **Application Tasks (A)**:
   - Grouping and Visualization of Class C GPCRs family types 232/1,510 sequences (C)
   - Grouping and Visualization of mGlu Class C subtypes (mGluR)
   - Analysis and Visualization of error classification (Error)
2. **Processed Data (D)**:

  - Full sequence (Full-Seq)
  - Extracellular domain (EC-Dom)
  - Transmembrane domain (TM-Dom)

3. **Data Transformations (T)**:
    - MSA
    - AAC
    - ACC
    - Digram (DI)
4. **Applied Methods (M)**:
    - GTM
    - KGTM
    - PT
    - Entropy-Based Measures (EBM)
    - Distribution Consistency (DC)
    - DSC (Distance Consistency)

1.7. Research Projects and Partners

Much of the work developed in this Thesis has been possible through the participation in several publicly-funded research projects, which are listed below:

- **Title:** Knowledge Acquisition in Pharmacoproteomics using Advanced Artificial Intelligence Methods (KAPPA AIM)
  Programme: Ministerio de Economía y Competitividad
  Project Reference: TIN2012-31377

Partners: Universitat Politècnica de Catalunya (UPC)
Main researcher: Alfredo Vellido (UPC)

- **Title:** Modelización matemática de las interacciones alostéricas complejas de los receptores acoplados a proteinas G: aproximaciones mecanísticas y probabilísticas.
  Programme: Ministerio de Economia, Industria y Competitividad - Área de gestión de Biomedicina
  Project Reference: SAF2010-19257
  Partners: Universitat Autònoma de Barcelona (UAB)
  Main researcher: Jesús Giraldo (UAB)

- **Title:** Validación de mGlu4 como diana terapéutica para el tratamiento multipotencial de las lesiones medulares.
  Programme: Fundació La Marató de TV3
  Project Reference: 110230
  Partners: Universitat Autònoma de Barcelona (UAB), Institut de Química Avançada de Catalunya (IQAC-CSIC), Institut de Génomique Fonctionnelle (IGF), Universitat de Montpellier
  Main researcher: Jesús Giraldo (UAB)

- **Title:** Integrated mathematical, computational and biochemical investigation of the crosstalk between metabotropic glutamate receptor 5 and dopamine D2 receptor: Relevance for the treatment of schizofrenia.
  Programme: Ministerio de Economia y Competitividad
  Project Reference: SAF2014-58396-R
  Partners: Universitat Autònoma de Barcelona (UAB)
  Main researcher: Jesús Giraldo (UAB)

- **Title:** Deciphering the role of peripheral and central nervous system metabotropic glutamate receptors in neuropathic pain with photoactivable ligands.

  Programme: ERA-NET NEURON Call for transnational research projects 2012

  Project Reference: PCIN-2013-018-C03-02. Includes 5 subprojects.

  Partners: Universitat Autònoma de Barcelona (UAB), Institut de Química Avançada de Catalunya (IQAC-CSIC), Institut de Génomique Fonctionnelle (IGF), CNRS, Universitat de Montpellier, INSERM, Montpellier, Universitat de Barcelona (UB), IRCCS Neuromed, University La Sapienza, Rome.

  Main researcher: Jesús Giraldo (UAB), from subproject 2 (SP2)

Part I

MATERIALS AND METHODS

Chapter 2

GPCR Data Collection and Transformation Methods

This Chapter summarily describes the materials employed in this Thesis, that is, the GPCR dataset acquired for our experiments. It also describes the several sequence transformations that were used to accommodate the symbolic sequences to the analytical methods.

2.1. Data Collection

The analyzed data were acquired from a GPCR-specific curated information repository, the GPCRdb [1], an enterprise started in 1993, which is part of the GLISTEN EU COST Action for the creation of a pan-European multidisciplinary research network [199], [83], [138].

[1] `http://gpcrdb.org`

The database divides the GPCR superfamily into five major families (A to E) based on the ligand types, functions, and sequence similarities (summarized in table 1.1). Within the families, proteins are further divided into groups (types and subtypes) which bind common agents on the extracellular side of the membrane.

The acquired set consists of non-redundant primary data: amino acid sequences in FASTA [121] format. Each position in a sequence is called a residue, which in turn, and as mentioned in previous Chapters, may be one of 20 possible AAs in a symbolic alphabet represented by a standard one-letter code. A sequence is therefore represented by an ordered combination of these letters. Table 2.1 shows the nomenclature for each amino acid and also two examples of sequences are shown in table 2.2.

Table 2.1: List of the 20 possible amino acids (AAs) in the GPCR sequence.

Amino acid name	Letter	Amino acid name	Letter
Alanine	A	Leucine	L
Arginine	R	Lysine	K
Asparagine	N	Methionine	M
Aspartate	D	Phenylalanine	F
Cysteine	C	Proline	P
Glutamate	E	Serine	S
Glutamine	Q	Threonine	T
Glycine	G	Tryptophan	W
Histidine	H	Tyrosine	Y
Isoleucine	I	Valine	V

Table 2.2: Two sequences from the dataset, shown for illustration. The first column represents the ID or header of the sequence and the second one represents the inner sequence. The gaps are represented by '−'.

Header	Sequence
ts1r3_mouse	RPKFLAWGEPVVLSLLLLLCLVLGLALAALGLSLVQA SGGSQFCFGLICLGLFCLSVLFPGRPSSASCLAQQPM AHLPLTGCLSTLFLQAAETFVESELPLSWNWLCSYLR GLWAWLVVLLATFVEAALCAWYLIAFPPEVVTDWSLP TEVLEHCHVRSLGLVHITNAMLAFLCFLGTFLVQSQP YNRARGLTFAMLAYFITWVSFVPLLANVQVAYCALGI LVTFHLPKCYVLLWLPKLNTQEFFLGRNAKK
q7pfp4_anoga	−FAFYTVVILSLIGIGISVLFLGLNLRF− − − − −ST ITVCGCMLVYTATILLGLDHSTL− − − − − −STICMRIY FLSAGFSLAFGSMFAKTFRVYRIFTH− − − − −LISVIG ALLLVDAFVVSFWMAAD− −C− − −WLG MLYAYKGLLLLVGVYMAWQTRNVK− −NDSQ YIGISVYSV VITSASVVVLANLLYERIITAG FVLISTTATLCLLFLPKI− − − − − − − − − − − − −

This Thesis focuses on class C GPCRs. Seven types of sequences belonging to this class, summarized in table 2.3, were investigated, namely: Metabotropic glutamate, Calcium sensing, GABA-B, Vomeronasal, Pheromone, Odorant and Taste.

Table 2.3: GPCR class C types.

GPCR Family C	Description
Type 1	Metabotropic glutamate
Type 2	Calcium sensing
Type 4	GABA-B
Type 5	Vomeronasal
Type 6	Pheromone
Type 7	Odorant
Type 8	Taste

Moreover, were investigated eight types of mGluRs, which belong to the first group of the GPCR family C (Type 1), namely mGluR1, mGluR2, mGluR3, mGluR4, mGluR5, mGluR6, mGluR7 and mGluR8. They are grouped into three groups summarized in table 2.4, based on their sequence, localization and signaling pathways.

Table 2.4: GPCR mGluR types.

GPCR mGluR Groups	Types
Group I	mGluR1, mGluR5
Group II	mGluR2, mGluR3
Group III	mGluR4, mGluR6, mGluR7, mGluR8

All in all, 5 data sets obtained from GPCRdb were analyzed at different stages in this Thesis:

- Data set 1 consists of 232 GPCRs sequences belonging to class C, which are further subdivided into 7 types: Metabotropic glutamate, Calcium sensing, GABA-B, Vomeronasal, Pheromone, Odorant and Taste. Type 3 was excluded from analysis of family C as it was not available in GPCRdb for the extracted data set. It consists of 76 mGlu, 9 CS, 45 GB, 8 VN, 42 Ph, 12 Od and 40 Ta receptors. The lengths of these sequences varied from 250 to 1,995 AAs.
- Data set 2 consists of 1,510 GPCRs sequences belonging to the previously listed 7 subtypes of class C. It consists of 351 mGlu, 48 CS, 208 GB, 344 VN, 392 Ph, 102 Od and 65 Ta receptors. The lengths of these sequences varied from 250 to 1,995 AAs.
- Data set 3 consists of 76 mGluR sequences included in data set 1, in turn sub-divided into 8 subtypes (mGluR1 to mGluR8) plus a group of mGluR-like sequences. They are distributed as 8 cases of mGluR1, 8 mGluR2, 11 mGluR3, 8 mGluR4, 11 mGluR5, 5 mGluR6, 10 mGluR8 and 15 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.

- Data set 4 consists of 351 mGluR sequences included in data set 2, 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 again be grouped into 3 categories: group I, group II and group III.
- Data set 5 consists of a subset of 1,252 GPCRs sequences from the original 1,510 that belong to class C and which includes an extracellular N-terminal domain description. These are further subdivided into the already described 7 subtypes. It consists of 282 mGlu, 45 CS, 156 GB, 293 VN, 333 Ph, 80 Od and 63 Ta receptors. The lengths of these sequences varied from 250 to 1,995 AAs.

2.2. Data Transformations

There is no biologically-relevant manner of representing the symbolic sequences describing proteins using real-valued vectors directly, but there are many principled sequence-transformation methods that make sequence analysis possible. In that sense, ML-based techniques require fixed-length vectors for training. However, protein sequences often have different lengths.

In this thesis, GPCR primary sequences have been transformed for their subsequent visualization analysis using DR methods. Three existing alignment-free transformations were used to limit the loss of information, the amino acid composition transformation (AAC), the auto cross covariance transformation (ACC) and the digram transformation (2-gram), which are described below.

Alternatively, the MSA method was applied to allow the application of conventional quantitative analysis techniques, but at the price of risking the loss of relevant information.

2.2.1. Alignment-free transformations

- The amino acid composition (AAC) transformation [171] uses the full-lenght of unaligned sequences. It consists on calculating the frequencies of the 20 amino acids of the sequence *alphabet* (i.e., a N x 20 matrix is obtained, where N is the number of items in the data set). As such, it ignores the sequential information itself (i.e., the relative position of the amino acids). Despite this, its use has previously yielded surprisingly solid results [171, 27].
- The Auto Cross Covariance (ACC) transformation [123], [41] is introduced to transform protein sequences into fixed-length vectors. Since each residue has many physical-chemical properties, such as hydrophobicity, hydrophilicity, normalized van der Waals volume, polarity, polarizability, sequence profile, etc., a sequence can be represented as a numeric matrix. For this, each sequence is first translated into physico-chemical descriptions by representing each amino acid with the five z-scales derived in [171], 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. This transformation generates an N x (z^2) matrix, where $z = 5$ is the number of descriptors. The maximal lag that was used for the ACC transforma-

tion is $l = 13$, which was found in previous studies to provide the best accuracy for this data set [33], [114].

- The digram-frequency transformation (2-gram) is a particular instance of the more general n-gram transformations. These transformations partially disregard sequential information to reflect only the relative frequency of appearance of AA subsequences (i.e., AAC is 1-gram). In the case of the Digram (2-gram) method, we calculate the frequencies of occurrence of each of the 400 possible AA pair combinations from the AA alphabet. Thus, this particular transformation generated an N x 400 matrix.

2.2.2. Sequences Alignment Transformation

Multiple Sequence Alignment transformation (MSA), is generally defined as the alignment of (usually) many biological sequences (protein or nucleic acid) of differing lengths. This method encodes structural information in similar protein sequences and reveals information about the GPCR structure. Then, structural similarity scores of the aligned sequences can be used in dimensionality reduction methods.

Many sophisticated MSA algorithms have been described in the literature. However, choosing the most suitable one for each dataset is by no means a trivial task. The characteristics of the sequences to be aligned, such as the shared identity, as well as their number and length, are aspects that have to be assessed in every MSA-based analysis.

Clustal Omega was the algorithm applied in our study because it is suitable to the analyzed dataset, which consists of sequences with a large N-Terminal domain. There are indeed alternative algorithms which could be applied for the alignment of the analyzed GPCRs that are worth mentioning, such as **TCoffee**

[143] and **MAFFT** with L-INS-i (iterative refinement with consistency from local pairwise alignment) [92], [93]. Both are recommended for their accuracy in progressive MSA applied to large datasets. Both have the drawback, though, that once errors are introduced at an early step of the alignment, they cannot be removed later [152], [87]. The T-Coffee algorithm uses a tree-based consistency objective function for alignment evaluation and produces an alignment by combining the output of several alignment methods. The MAFFT algorithm, in turn, introduces the fast Fourier tranformation (FFT) in sequence alignment in which an amino acid sequence is transformed into a sequence composed of volume and polarity values for each residue. Note, in any case, that this thesis is more focused on the use of unaligned data transformations and that we have consciously limited our research on alternative MSA methods.

The similarity between two sequences is evaluated by first aligning the sequences (or parts of them) and then deciding whether their alignment is more likely to have occurred because the sequences are related or just by chance. When two sequences are compared, the basic mutational processes under consideration are *substitutions*, which change residues in a sequence, and *insertions* and *deletions*, which add or remove amino acids in the sequence. Insertions and deletions are together referred to as *gaps*. Then, the score, used to judge the correctness of the alignment, is modified accordingly to allow the number of gaps to be limited.

The total similarity score assigned to an alignment will be a sum of terms for each aligned pair of residues, plus terms for each gap. In a probabilistic interpretation, this corresponds to the logarithm of the relative likelihood that the sequences are related, compared to being unrelated. Thus, all the scores are arranged in a 20 x 20 matrix known as *score matrix* or *substitution matrix*, consisting on arrays of symbols from the 20 amino acid alphabet [43].

Chapter 3

Methods

This Chapter provides, first, an overview of different machine learning techniques of the manifold learning family that are used as methods for simultaneous data visualization and grouping. This is followed by a brief description of the phylogenetic trees used for GPCR grouping structure investigation, in the pursuit of better interpretability of the results.

3.1. Dimensionality Reduction Techniques for Exploratory Data Visualization

Visualization is used in this Thesis as an exploratory Data Mining tool, facilitating us to veer from a strictly deductive mode of research towards an inductive approach to knowledge discovery. That is, we aim to generate a faithful visualization of the available Multivariate Data (MVD) in the hope that it will provide us with non-trivial clues regarding data structure that might lead to hypothesis generation [94], [196]. By means of ML approaches, visualization can proof to be extremely informative in domains in which data structure is not fully known or is uncertain.

The visualization of MVD involves, in one way or another, a process of data dimensionality reduction. This is a very general problem in pattern recognition at large and ML in particular for whose solution a broad palette of approaches and methods have been proposed. Covering them is of course beyond the scope of this Thesis and, therefore, we will focus on techniques of the manifold learning family and associated kernel-based methods.

3.1.1. The basic GTM

Generative Topographic Mapping (GTM) [19] is a non-linear latent variable model of the manifold learning family, with sound foundations in probability theory. It performs simultaneous clustering and visualization of the observed MVD through a topology-preserving mapping from a latent space in $\mathbb{R}^{\mathbb{L}}$ (with L being usually 1 or 2 for data visualization purposes) onto the $\mathbb{R}^{\mathbb{D}}$ space in which the observed MVD reside. The mapping that generates the embedded manifold is functionally described as:

$$y = W\phi(u), \tag{3.1}$$

where u is an L-dimensional point in latent space, W is the matrix that generates the mapping, and ϕ consists of S basis functions ϕ_s(radially symmetric Gaussians in the standard model for continuous data). To achieve computational tractability, the prior distribution of u in latent space is constrained to form a uniform discrete grid of M centres in the form of a sum of *delta* functions:

$$p\left(u\right)=\frac{1}{M}\sum_{m=1}^{M}\delta\left(u-u_{m}\right), \tag{3.2}$$

where M is the number of nodes in the grid.

This way defined, the GTM can also be understood as a special case of a Gaussian mixture model that is adapted to provide MVD visualization. Each component m in the mixture defines the probability of an observable data point $\mathbf{x}$ given a latent point u_m and model:

$$p(x \mid u_m, \Theta)=\left(\frac{\beta}{2\pi}\right)^{\frac{D}{2}}\exp\left\{-\frac{\beta}{2}\left\|x-y_m\right\|^2\right\} \tag{3.3}$$

where D is the dimensionality of the data space, and $y_m = W\phi\left(u_m\right)$.

The adaptive parameters Θ include W and the common inverse variance β. A density model in data space is therefore generated for each component m of the mixture, which, assuming that the observed MVD X consists of N independent, identically distributed (i.i.d.) data points x_n, leads to the definition of a likelihood in the form:

$$\mathcal{L}\left(W,\beta\right)=\prod_{n=1}^{N}\frac{1}{M}\sum_{m=1}^{M}p\left(x_n \mid u_m, W, \beta\right) \tag{3.4}$$

However, it is more convenient to work with the log-likelihood function:

$$L\left(W,\beta\right)=\sum_{n=1}^{N}\ln\left\{\frac{1}{M}\sum_{m=1}^{M}p\left(x_n \mid u_m, W, \beta\right)\right\} \tag{3.5}$$

The adaptive parameters of the model are usually optimized by Maximum Likelihood (M-L) using the Expectation-Maximization (EM) algorithm [40]. In the E-step, the current values of the parameters W and β are used to evaluate the posterior probability, or *responsibility*, that each component m takes for every data point x_n, which, using Bayes' theorem, is given by

$$R_{nm} \equiv p\left(m \mid x_n\right) = \frac{p\left(x_n \mid m\right)}{\sum_j p\left(x_n \mid j\right)}, \tag{3.6}$$

in which the prior probabilities $P\left(m\right) = \frac{1}{K}$ have cancelled between numerator and denominator. Using 3.3, we can rewrite this in the form

$$R_{nm} = \frac{exp\left\{-\frac{\beta}{2} \parallel x_n - y_m \parallel^2\right\}}{\sum_m exp\left\{-\frac{\beta}{2} \parallel x_n - y_m \parallel^2\right\}} \tag{3.7}$$

In the M-step of the algorithm we then use these responsibilities to re-estimate the weight matrix W by solving the following system of linear equations:

$$\left(\Phi^T G \Phi\right) W_{new}^T = \Phi^T R X, \tag{3.8}$$

which follow by maximization of the expected complete-data log likelihood. In 3.8, Φ is a $K \times M$ matrix with elements $\Phi_{mj} = \Phi_j\left(u_m\right)$, X is an $N \times D$ matrix with elements x_{nk}, R is a $K \times N$ matrix with elements R_{nm}, and G is a $K \times K$ diagonal matrix with elements. The inverse variance parameter is also re-estimated in the M-step:

$$\frac{1}{\beta_{new}} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{m=1}^{M} R_{nm} \parallel W_{new} \phi\left(u_m\right) - x_n \parallel^2 \tag{3.9}$$

We can initialize the parameters W so that the GTM model approximates a PCA projection. To do this, we first evaluate the data covariance matrix and obtain the eigenvectors corresponding to the q largest eigenvalues, and then we determine W by minimizing the sum-of-squares error between the projections of the latent points into data space by the GTM model and the corresponding projections obtained from PCA. The value of β^{-1} is initialized to be the larger of either the $q + 1$ eigenvalue from PCA (representing the variance of the data away from the PCA sub-space) or the square of half of the grid spacing of the PCA-projected latent points in data space.

The main advantage of the GTM over the functionally similar Self-Organizing Map (SOM) algorithm (described below in some detail)is that the former generates a density distribution in the input data space so that the model can be described and developed within a principled probabilistic framework. An example of development of the GTM is the use of a Bayesian approach to automatic regularization and smoothing of the resulting mapping. As part of this process, the GTM learning parameters calculation is grounded in a sound theoretical basis. The GTM also provides the well-defined objective function of equation 3.5, whereas the SOM training does not involve the minimisation of any error function; its maximisation using either standard techniques for non-linear optimisation or the EM-algorithm has been proved to converge, unlike in the case of the SOM.

3.1.2. Foundations of Kernel Dimensionality Reduction Models

Generally speaking and as applied to our research, the purpose of using an unsupervised kernel learning method for the analysis of protein sequence data is finding a group of GPCRs such that similar sequences belong to the same group, in a way that we are able to find a group of sequences such that similarities are much greater than the similarities among sequences from different groups.

Unsupervised methods that were capable of providing simultaneous grouping and visualization of sequence data would be especially adequate for this type of problems, as visualization can help us to intuitively interpret the grouping and classification results by providing intuitive insights about the relationships between groups. The visualization of the high-dimensional GPCR sequences would considerably help to understand their global grouping structure.

Most DR strategies, though, have been designed for real-valued data. Needless to say, protein symbolic sequences of amino acids do not fit into this description, and alternative strategies are thus required. Over the last few years, several kernel methods for the visualization (and eventually clustering) of non-standard multivariate data have been proposed. The use of kernels allows mapping data implicitly into a high-dimensional space called feature space, in such a way that computing a linear partitioning in this feature space results in a corresponding non-linear partitioning in the observed data space.

In this section, we describe the basis of some methods that we consider to be representative of the current available choices in the field and which should help to lay the conceptual foundations of the kernel manifold learning models used in the Thesis.

Kernel Principal Component Analysis

Principal Component Analysis (PCA) [154] is an orthogonal transformation of the coordinate system in which we describe the observed MVD. The central idea of PCA is to achieve dimensionality reduction while retaining as much of the variation present in the data set as possible. Dimensionality reduction is achieved because a small number of principal components often suffices to account for most of the variance (structure) in the data.

Data are effectively transformed by projecting them into the subspace spanned by the first k eigenvectors of the covariance matrix of the analyzed data set. The new coordinates are known as the principal coordinates with the eigenvectors referred to as the principal axes. Details of this technique can be found elsewhere [86].

Kernel PCA [175], or KPCA, is the application of PCA in a kernel-defined feature space making use of the dual representation. This method makes possible to detect non-linear relations between variables in the data by embedding the data into a kernel-induced feature space, where linear relations can be found by means of PCA. Also, KPCA can be seen as a way of inferring a low-dimensional explicit geometric feature space that best captures the structure of the data.

The projection of a new data point $\phi\left(x\right)$ onto the direction u_j in the feature space, is given by

$$\begin{aligned} P_{u_j}\left(\phi\left(x\right)\right) = u_j^{'}\phi\left(x\right) &= \left\langle \sum_{i=1}^{l} \alpha_i^j \phi\left(x_i\right), \phi\left(x\right) \right\rangle \qquad (3.10) \\ &= \sum_{i=1}^{l} \alpha_i^j \left\langle \phi\left(x_i\right), \phi\left(x\right) \right\rangle = \sum_{i=1}^{l} \alpha_i^j K\left(x_i, x\right) \qquad (3.11) \end{aligned}$$

Hence, we will be able to project new data onto the eigenvectors in the feature space by performing an eigen-decomposition of the kernel matrix.

Let be U_k the subspace spanned by the first k eigenvectors in the feature space. Then, we can compute the k-dimensional vector projection of new data into this subspace as

$$P_{U_k}\left(\phi\left(x\right)\right) = \left(u_j^{\prime}\phi_j\left(x\right)\right)_{j=1}^{k} = \left(\sum_{i=1}^{l} \alpha_i^j K\left(x_i, x\right)\right)_{j=1}^{k} \qquad (3.12)$$

where $\alpha^j = \lambda_j^{-\frac{1}{2}} v_j$ is given in terms of the corresponding eigenvector λ_j and eigenvalue v_j of the kernel matrix. Equation 3.12 forms the basis of KPCA.

The critical question for assessing the performance of KPCA is the extent to which the projection captures new data drawn according to the same distribution as the training data. Therefore, we assess the stability of KPCA through the pattern function:

$$f(x) = \parallel P_{U_k}^{\perp}(\phi(x)) \parallel^2 = \parallel \phi(x) - P_{U_k}(\phi(x)) \parallel^2 = \\ \parallel \phi(x) \parallel^2 - \parallel P_{U_k}(\phi(x)) \parallel^2$$

That is, the squared norm of the orthogonal (residual) projection for the subspace U_k spanned by the first k eigenvectors. As always we wish the expected value of the pattern function to be small

$$E_X[f(x)] = E_X\left[\parallel P_{U_k}^{\perp}(\phi(x)) \parallel^2\right] \approx 0$$

Thus, capturing a high proportion of the data variance in an small number of dimensions is an indication that a reliable set of features has been detected and that the corresponding subspace will capture most of the variance of yet unobserved test data.

Kernel Self-Organizing Maps

KPCA provides a method according to which we can visualize GPCR sequences in a representation space (e.g. spanning only two PCs). Unfortunately, this visualization through projection is not accompanied by a grouping or clustering of the sequences. The Self Organising Maps (SOM), also popularly referred to as Kohonen network [103], [102] is a computational intelligence (CI) method for the visualization of high-dimensional data that also provides vector quantization and, in doing so, allows the partition of the data into clusters.

The SOM defines a topologically-ordered mapping that generates the projection of observed multivariate data items onto a regular, usually two-dimensional map. This map consists of a regular lattice of processing units, also called *neurons* (due to the original description of SOM as a bio-plausible model of cognitive processes). Each of these units is associated to a prototype vector in the observed data space, which can be considered as a representative example of a given subset of data cases. The map attempts to represent all the available data cases with optimal accuracy using a restricted set of prototypes. Each prototype could therefore be understood as a cluster representative.

The resulting map is meant to retain the topological order of the observed space, so that similar prototypes in the observed space are also close to each other in the visualization map.

In its standard form, the SOM algorithm distinguishes two stages: the competitive stage and the cooperative stage. In the former, the SOM *neuron* best matching a given data case is selected, while, in the latter, the coefficients (or *weights*) of the best-matching prototype (and to a lesser extent, those of its immediate lattice neighbors) are changed to become fractionally closer to that data case.

More formally, let $X = [x_1, x_2, ...x_d]^T \in R^d$ be the input vector. Assume a discrete lattice of units indexed with a index i. Each unit is associated to a corresponding weight vector (prototype) $W = [w_1, w_2, ...w_d]^T \in R^d$. Data case X_n is mapped to that unit whose weight vector is its nearest neighbour, from among all the weight vectors. This is called the best-matching unit (BMU) and is found as: $BMU_n = argmin_i \|X_n - W_i\|$

Thus, the training process of the SOM algorithm can be summarized as follows:

- For each observed data case, find out the nearest-neighbour (winner) from among the weight vectors associated to the map.
- Update the weights of the winner and all its neighbours according to some updating criterion.
- Iterate the process for all data cases (in an online or batch procedure) until some convergence criterion is met.

The SOM model, though, has some limitations due to its heuristic nature. In summary:

- Different runs of the SOM algorithm with different initializations yield different results.
- The selection of its parameters (e.g., learning rate, or neighbourhood function type or size) has no theoretical basis.
- There is no guarantee of error convergence for the training procedure. Neighbourhood preservation is not guaranteed either.
- There is no theoretical basis for complexity control (regularization and overfitting)

Furthermore, the Euclidean distance used to describe similarity in the standard SOM model is not adequate for the analysis of non-real-valued data such as symbolic protein sequences.

A kernel version of the SOM, namely the Kernel Self-Organizing Map, or KSOM, was proposed by MacDonald and Fyfe [125]. It can be understood as a kernelization of the k-means clustering algorithm, but with added neighbourhood learning. More precisely, a kernel function is applied to transform the input (observed data) into a high-dimensional feature space, thus transforming the distance metric to nonlinear and adding more flexibility in the vector-quantization process in order to better capture the data structure [115]. Each data case x is mapped to the feature space via a nonlinear function $\phi(x)$. In principle each mean can be described as a weighted sum of the observations in the feature space,

$$m_i = \sum_n \gamma_{i,n} \phi(x_n)$$

where $\{\gamma_{i,n}\}$ are the constructing coefficients. The algorithm then selects a mean or assigns a data case with the minimum distance between the mapped point and the mean,

$$\|\phi(x) - m_i\|^2 = \| \phi(x) - \sum_n \gamma_{i,n} \phi(x_n) \|^2 = \tag{3.13}$$

$$K(x,x) - 2\sum \gamma_{i,n} K(x,x_n) + \sum_{n,m} \gamma_{m,n} K(x_n, x_m) \tag{3.14}$$

The update of the mean is based on an update expression similar to that of the SOM:

$$m_i(t+1) = m_i(t) + \Lambda[\phi(x) - m_i(t)] \tag{3.15}$$

where Λ is the normalized winning frequency of the i-th mean, defined as:

$$\Lambda = \frac{\xi_{i(x),j}}{\sum_{n=1}^{t+1} \xi_{i,n}} \tag{3.16}$$

and ξ is the winning counter and is often defined as a Gaussian function between the indexes of the two neurons. As the mapping function ϕ is not known, the updating rule 3.15 is further elaborated and leads to the following updating rules for the constructing coefficients of the means [125]:

$$\gamma_{i,n}(t+1) = \begin{cases} \gamma_{i,n}(t)(1-\xi), & for\ n \neq t+1 \\ \xi, & for\ n = t+1 \end{cases}$$

Note that these constructing coefficients, $\gamma_{i,n}$, together with the kernel function, effectively define the kernel SOM in the feature space. The winner selection, i.e. 3.13, operates on these coefficients and the kernel function. No explicit mapping function ϕ is required. The exact means or neuron weights m_i, are not required [211].

There is an alternative direct way to kernelize the SOM by mapping the data points and neuron weights, both defined in the input space, to a feature space; this is followed by applying standard SOM in the mapped dot-product space. The winning rules of this second type of KSOM have been proposed as follows, either in the input space [153], $v = arg\min_i \|x - m_i\|$ or in the feature space [5], $v = arg\min_i \|\phi(x) - \phi(m_i)\|$

These two rules are equivalent for certain kernels, such as the Gaussian. The weight update rule proposed in [5] is:

$$m_i(t+1) = m_i(t) + \alpha(t)\,\eta(v(x), i)\,\nabla J(x, m_i) \tag{3.17}$$

where $\nabla J(x, m_i) = \|\phi(x) - \phi(m_i)\|^2$ is the distance function in the feature space or the proposed instantaneous or sample objective function. Also, $\alpha\left(t\right)$ and $\eta\left(v\left(x\right),i\right)$ are, in turn, the learning rate and neighbourhood function.

Note that

$$J\left(x,m_i\right)=\|\phi\left(x\right)-\phi\left(m_i\right)\|^2=K\left(x,x\right)+K\left(m_i,m_i\right)-2K\left(x,m_i\right)$$

and,

$$\nabla J\left(x,m_i\right)=\frac{\partial K\left(m_i,m_i\right)}{\partial m_i}-2\frac{\partial K\left(x,m_i\right)}{\partial m_i}$$

Therefore this kernel SOM can also be operated entirely in the feature space with the kernel function. As the weights of the neurons are defined in the input space, they can be explicitly resolved.

The standard SOM minimizes the following energy function [113], [75]:

$$E=\sum_i\int_{V_i}\sum_j\eta\left(i,j\right)\parallel x-m_j\|^2p\left(x\right)dx$$

where V_i is the Voronoi region of neuron i.

The extension of this energy function in the feature space is:

$$E_F=\sum_i\int_{V_i}\sum_j\eta\left(i,j\right)\parallel\phi\left(x\right)-\phi\left(m_j\right)\|^2p\left(x\right)dx$$

The KSOM can be seen as a result of directly minimizing this transformed energy. Using the sample gradient on $\eta\left(v\left(x\right),j\right)\parallel\phi\left(x\right)-\phi\left(m_j\right)\|^2$, we obtain:

$$\frac{\partial \hat{E_F}}{\partial m_i} = \frac{\partial}{\partial m_j} \sum_j \eta \left(v\left(x\right), j\right) \parallel \phi\left(x\right) - \phi\left(m_j\right) \parallel^2 = -2\eta\left(v\left(x\right), i\right) \nabla J\left(x, m_j\right),$$

which leads to the same weight update expression for the KSOM as in equation 3.17.

Although KSOM makes the standard Kohonen map much more flexible, it still inherits the limitations of SOM outlined above. The analysis of GPCR sequences would benefit from a model with solid grounds on probability theory that might benefit from the automatic optimization of all its parameters. One such kernel model of the manifold learning family is proposed and applied to the analysis of GPCR sequences in the following Chapter.

3.2. Kernel GTM

Kernelization of the GTM

Kernelization is a method originally defined for Support Vector Machines (SVM) that could be used to develop generalizations of any algorithm that could be cast in terms of a mathematical dot product. The basic premise is that a method formulated in terms of kernels can use the one that best suits the problem and data type at hand. With this purpose, we here define kernel-GTM (KGTM). It takes advantage of the original GTM functionalities and, in particular, to achieve a simultaneous clustering and visualization of a wide variety of data types [146]. The rationale for this extension is that the original standard GTM lacks the ability to handle more structured data, such as the strings of symbols of the protein primary sequences.

The KGTM model as applied to sequence analysis

Let us consider the problem of embedding GPCR sequences in a high-dimensional space in such a way that their relative position in that space reflects their similarity and that the inner product between their images in that space can be computed efficiently. The first decision to be made is what notion of similarity should be reflected in the embedding, or, in other words, what features of the symbolic sequences are informative for such a task.

The meaning of similarity in biological applications can be related to both functional similarity and symbolic sequence similarity, the latter being measured by the number of insertions, deletions and symbol replacements in the sequence. Measuring sequence similarity should therefore provide us with a good indicator of the functional similarity that we would like to capture.

The similarity between two sequences is usually evaluated by first aligning the sequences (or parts of them) and then deciding whether their alignment is more likely to have occurred either because the sequences are related, or just by chance.

When two sequences are compared, the basic mutational processes under consideration are *substitutions*, which change residues (amino acids) in a sequence, and *insertions* and *deletions*, which add or remove amino acids in the sequence. Insertions and deletions are together referred to as *gaps*. Natural selection has an effect on this process by screening the mutations, so that some types of changes remain throughout evolution and appear more often than others [43].

In order to have some control over the number of gaps, their size, position, etc., gap *penalties* are usually introduced. The score, used to judge the correctness of the alignment, is then modified accordingly to allow the number of gaps to be limited.

The total similarity score assigned to an alignment will be a sum of terms for each aligned pair of residues, plus terms for each gap. In a probabilistic interpretation, this corresponds to the logarithm of the relative likelihood that the sequences are related, compared to being unrelated. Informally, identities and conservative substitutions are expected to be more likely in alignments than appearing by chance and, therefore, contribute positively to the similarity score. On the contrary, non-conservative changes are expected to be observed less frequently in real alignments than expected to happen by chance, and so they contribute negatively to the score.

In order to gauge similarity for each aligned residue pair, we will derive substitution scores from our probabilistic model. The scores can be arranged in a matrix. For the protein sequences analyzed in our research, consisting on arrays of symbols from a 20 amino acid *alphabet*, a 20 × 20 matrix can be calculated, known as *score matrix* or *substitution matrix*.

A kernel function can be thought of as a measure of similarity between sequences. Different kernels correspond to different notions of similarity, and can lead to discriminative functions with different performance. The kernel function designed to analyze GPCRs with KGTM is a variation on that described in [146], based on the mutations and gaps between sequences:

$$K\left(x,x'\right) = exp\left\{\nu \frac{\pi\left(x,x'\right)}{\sqrt{\pi\left(x,x\right)\pi\left(x',x'\right)}}\right\} \tag{3.18}$$

where x and x' are two sequences and ν is a prefixed parameter; $\pi\left(.\right)$ is a score function commonly used in bioinformatics and expressed as: $\pi\left(x,x'\right) = \sum_r s\left(x_r,x'_r\right) - \gamma$, where x_r and x'_r are the r^{th} residue in the sequences. The value of $s\left(x_r, x'_r\right)$ can be found in a mutation matrix [43] and γ is a gap penalty (usually the number of gaps in sequences). As a contribution, a normalization factor, defined as the geometric mean of the maximum scores for each of the sequences, is used in the kernel function instead of the sum used in [146].

The KGTM algorithm

The kernel trick allows the observed data X to be implicitly mapped onto a high-dimensional feature space H via a nonlinear function: $x \longmapsto \psi\left(x\right)$. A similarity measure can then be defined from the dot product in space H as follows:

$$K\left(x, x^{'}\right) = \left\langle \psi\left(x\right), \psi\left(x^{'}\right)\right\rangle \tag{3.19}$$

K is a kernel function that should satisfy Mercer's condition [176]. It allows us to deal with learning algorithms using linear algebra and analytic geometry. In general, this method deals with data in the high-dimensional dot product space H, usually known as feature space.

The use of kernel trick avoids the explicit estimation of ψ, whose dimension is usually unknown (or even infinite).

The kernelization of GTM can be implemented by redefining equation 3.3 in feature space as:

$$p\left(\psi\left(x\right) \mid u_m, \Theta\right) = \left(\frac{\beta}{2\pi}\right)^{\frac{D}{2}} \exp\left\{-\frac{\beta}{2}\|\psi\left(x\right) - y_m\|^2\right\} \tag{3.20}$$

Note that the prototypes y_m are now defined in the feature space and not in data space, as originally. In most cases, the term $\|\psi\left(x\right) - y_m\|^2$ cannot be

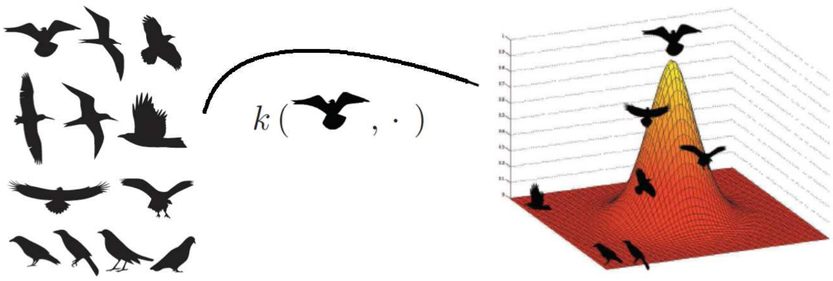


Figure 3.1: Example of kernel function as similarity measure between input objects, which computes the inner product into a feature space. It shows that data not linear separable in input space map into some feature space where data is linear separable.

directly evaluated, given that the function $\psi(\cdot)$ is usually unknown. However, this term can be also expressed as follows:

$$\|\psi(x) - y_m\|^2 = \langle\psi(x), \psi(x)\rangle + \langle y_m, y_m\rangle - 2\langle\psi(x), y_m\rangle \tag{3.21}$$

Here, we assume that, as in KPCA, y_m can be expanded on the training data in the feature space. That is, $y_m = \mathbf{\Psi} w_m$, where $\mathbf{\Psi}$ is a $D \times N$ -matrix of vector columns $\mathbf{\Psi}(x_n)$, $n = 1..N$, and w_m a weight vector. With the aim of preserving the topology, we correlate the weight vector to the latent space by $w_m = \Lambda\phi_m$, where Λ is an adaptive weight matrix and $\phi_m = \phi(u_m)$ is the set of radial basis functions typically used by GTM. Therefore, equation 3.21 becomes:

$$\|\psi(x) - y_m\|^2 = J_{mn} = K_{nn} + (\Lambda\phi_m)^T \mathbf{K}\Lambda\phi_m - 2k_n\Lambda\phi_m, \tag{3.22}$$

where $\mathbf{K}$ is a kernel matrix with elements $K_{nn'} = \langle\psi(x_n), \psi(x_{n'})\rangle$, and row vectors k_n . Thereby J_{mn} is expressed in terms of the kernel matrix, making the definition of function $\psi(\cdot)$ unnecessary. The adaptive parameters of the model are now Λ and β , which can be optimized by ML using EM, as in GTM. The likelihood of the model is formulated as follows:

$$\mathcal{L}(\Lambda, \beta) = \prod_{n=1}^{N} \frac{1}{M} \sum_{m=1}^{M} p\left(\psi\left(x_n\right) \mid u_m, \Lambda, \beta\right). \tag{3.23}$$

Following the usual EM algorithm, we are specially interested in one of the results of the expectation step of EM, namely the estimation of the posterior distribution $R_{mn} = p\left(u_m \mid \psi\left(x_n\right), \Lambda, \beta\right)$, defined as:

$$R_{mn} = \frac{p\left(\psi\left(x_n\right) \mid u_m, \Lambda, \beta\right)}{\sum_{m'=1}^{M} p\left(\psi\left(x_n\right) \mid u_{m'}, \Lambda, \beta\right)} \tag{3.24}$$

R_{mn} measures the degree of responsibility (probability) of a point u_m in the latent space for the generation of a $\psi\left(x_n\right)$ GPCR data subsequence. In turn, each R_{mn} is an element of a $M \times N$ responsibility matrix R.

In the maximization step we use equation 3.23 as the optimization function to determine the parameters Λ and β, which results in the following expressions:

$$\Lambda^T = \left(\Phi^T G \Phi\right)^{-1} \Phi^T R \tag{3.25}$$

$$\frac{1}{\beta} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{m=1}^{M} R_{mn} J_{mn} \tag{3.26}$$

The initial values for the parameters of KGTM are selected using KPCA (a procedure which is inspired in the PCA-based initialization of parameters for the standard GTM).

3.3. Phylogenetic Trees for Hierarchical Data Visualization

As mentioned in section 1.2, a phylogenetic tree (PT) is a dendrogram-like graphical representation of the evolutionary relationship between taxonomic groups. In biology, the term phylogeny refers to the evolution or historical development of a plant or animal species. Taxonomy is the system of classifying species by grouping them into categories according to their similarities in their physical or genetic characteristics. Phylogenies are useful for organizing knowledge of biological diversity, for structuring classifications, and for providing insight into events that occurred during evolution. Most Pts are rooted, meaning that one branch (which is usually unlabeled) corresponds to the common ancestor of all the species included in the tree. However, a tree can be drawn in any orientation [13]. PTs are not meant to be understood as completely true and accurate descriptions of the evolutionary paths they represent, because in any of them there are a number of possible evolutionary pathways that could produce the pattern of relatedness illustrated. More precisely, and in the case of protein sequences, they only illustrate the probability that two sequences are more closely related to each other than to a third one.

In this Thesis, PTs were visualised using two software tools for tree visualization, namely *Jalview* and *Treevolution.* Jalview 2.6.1, with the Blocks of Amino Acid Substitution Matrix 62 (BLOSUM62) [74], [44], which is the standard for most programs that use this type of matrices. In this application, sequences are introduced in FASTA format [121] and the trees are calculated on the basis of a measure of similarity between each pair of sequences in the alignment. Treevolution [1] [174] is a software developed in Java that integrates

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

the Processing package [11]. 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-guided highlighting of protein families helps the user to focus on sequence groupings of interest. The PT visualized in *Treevolution* is obtained using the software *Clustal Omega* [181]. This application, in which sequences data are introduced in FASTA format, performs the MSA [50] with the distance-based PT reconstruction method called neighbor-joining (NJ) [169]. The NJ method provides both the topology and the branch lengths of the final tree, which is again calculated on the basis of the BLOSUM62 scoring matrix. Then, a similarity measure based on the aligned data is given in the PT visualization.

For the experiments in Chapter 4, the MSA-based clustering method named Unweighted Pair-Group Method with Arithmetic Mean (UPGMA) [183], which also generates a standard PT, was applied to data set 1. Both distance based algorithms differ from each other. While UPGMA assumes the very rare condition that the molecular clock is perfect, using a rooted tree which means no variation in evolution rates across GPCRs, the NJ algorithm does not assume a perfect molecular clock, uses an unrooted tree and ensures that the clusters that are merged in the course of tree reconstruction are not only close to each other (as in UPGMA) but also far apart from the rest. Note, in any case, that we used PTs in this thesis as a counterpart for the data grouping algorithms we propose. For this reason, we are mostly interested in their behaviour as hierarchical grouping techniques, represented in the form of dendograms.

In summary, the PTs and the visualization approaches previously described in section 3.1 differ from each other. On the one hand, PTs adopts a hierarchical clustering approach from aligned versions of the sequences and only reflects relative similarity, whereas the latter do not reflect hierarchy but implicitly,

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

while reflecting similarity in projective form. These approaches, though, nicely complement each other.

Part II

EXPERIMENTS: SETTINGS, RESULTS AND DISCUSSION

Chapter 4

Grouping and visualization of Class C GPCRs family types

In this Chapter, we report our first set of experimental results, which concern the exploratory visualization and grouping of the different GPCR class C types using a kernel manifold learning technique, namely KGTM, as well as PTs.

4.1. Results and discussion

The visualization of the class C GPCR sequences is carried out here for data set 1 as described in section 2.1. These data were transformed using the MSA method described in section 2.2.2 and then fed to the KGTM model using the kernel defined by Eq. 3.18 and to the standard PT to generate exploratory visualizations.

4.1.1. Visualization of data set 1 using KGTM

The visualization results obtained using KGTM are shown in Fig.4.1. There is quite clear separation between many of the GPCR class C subtypes, which are shown in the latent space of the model using the *mode-projection*, defined as:

$$m_{mode} = \underset{m}{\texttt{argmax}}\, R_{mn}, \tag{4.1}$$

where R_{mn} is defined in Eq. 3.24.

Many of these subtypes occupy a rather differentiated area on the map, showing little overlapping. A few of them, though, have clearly overlapping representations. Both cases could be the source of insight on the peculiarities of subtype structure. *Metabotropic glutamate* (subtype 1), *GABA-B* (3), and *Taste* (7) are clearly differentiated from the rest of subtypes, which, in turn, show significant overlapping between them.

The *mode-projection* is an intuitive form of visualization that sacrifices detail in favour of clarity. By using only the maximum of the responsibilities in $\mathbf{R}$, though, it disposes of much of the rich information that might be contained in this matrix of probabilities.

There are different ways of visually representing this information. One of them is the display of *maps of probability* $\mathbf{R}_i$, for a given sequence i. Sequences clearly ascribed to a subtype are likely to have their responsibilities concentrated in only a few modes (latent points), whereas the probabilities of sequences with-

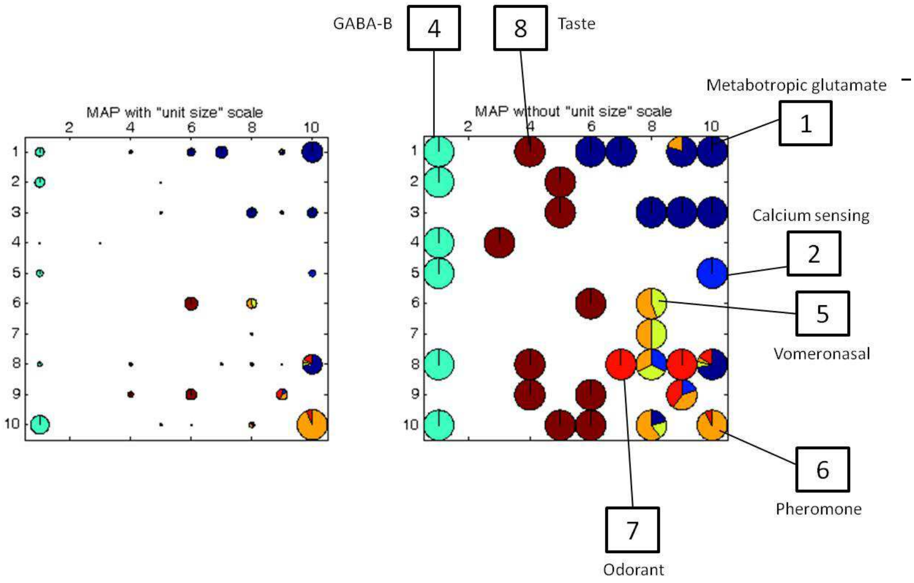
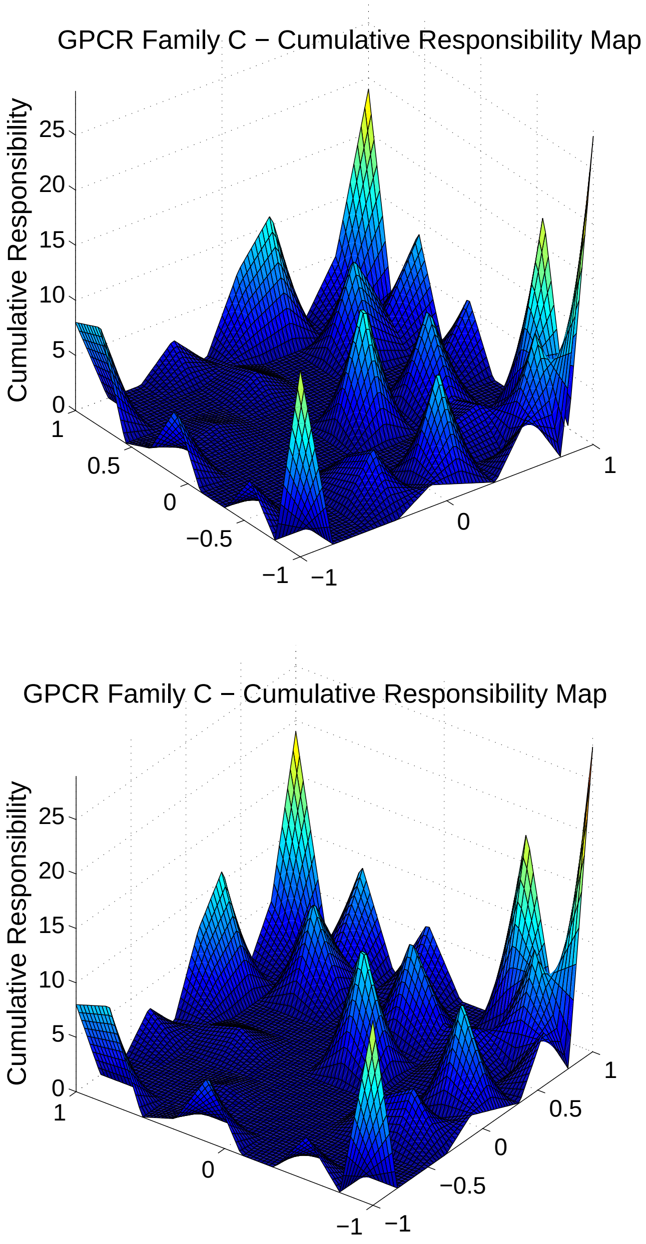


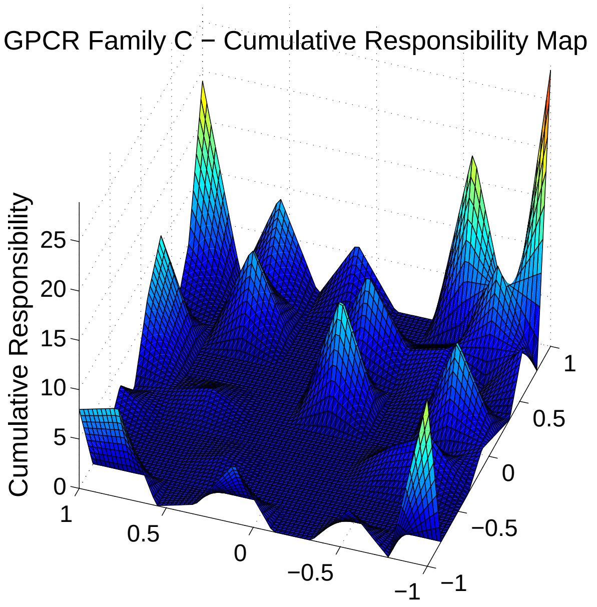
Figure 4.1: Data visualization on a 10 × 10 KGTM representation map, using the mode-projection as described in the text. Left) Each of the pie charts corresponds to a latent point, and their size is proportional to the ratio of sequences assigned to them. Each portion of a chart corresponds to the percentage of sequences belonging to each subclass, coded in different colours. Right) The same map is provided without sequence ratio-based size scaling, to ease the interpretation of the visualization. Labels as described in the text tags.

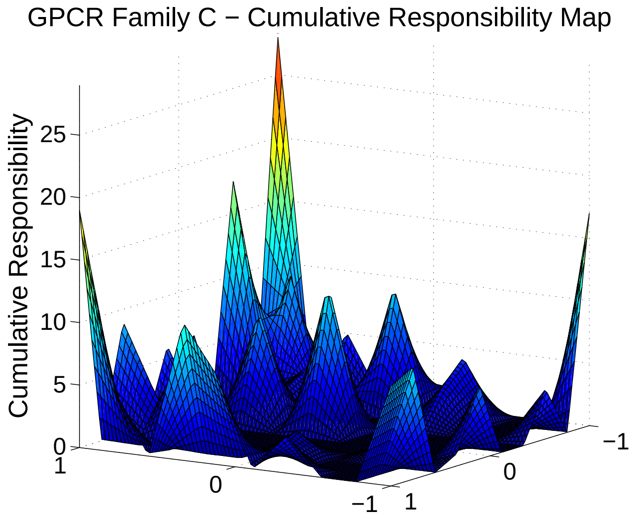
out clear subtype ascription may be more evenly spread across the map.

We may be also interested in the responsibilities of all sequences of a given subtype at once. In this case, we would aim to assess if each subtype has its responsibilities located in a well-defined area of the map or not. For this, we can use the cumulative responsibility of the sequences that belong to a given type c, which is defined as a vector $CR_c = \sum_{\{n \in c\}} R_{mn}$, for $m = \{1, .., M\}$.

Figure 4.2: Visualization of the global CR (on the vertical axis) of the data set on the representation map. For better appreciation, several viewpoints of the map are provided.





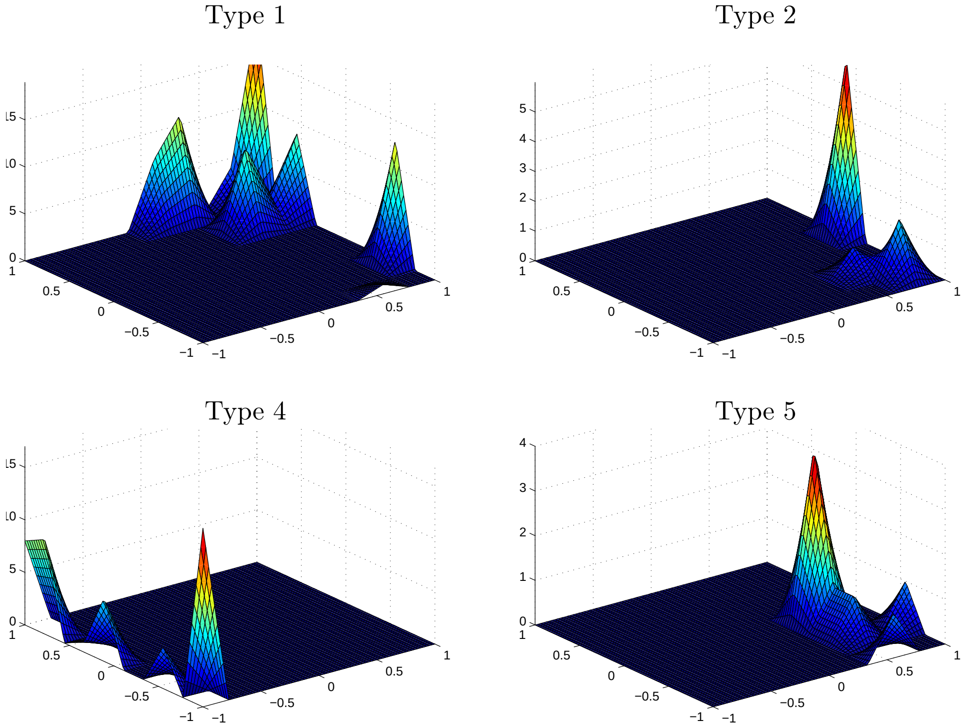


This takes us to the possibility of displaying the cumulative responsibility of all sequences in the data set. With this map of probability, the existence of

CR peaks and valleys can be explored. The latter are likely to define the model estimated boundaries between subtypes.

The global CR is displayed in figure 4.2, whereas figure 4.3 provides the visualization of the CR_c for the seven analysed subtypes of the class C. Consistent with the subtype specific representations in figure 4.3, several local maxima are shown to correspond to each type, which could be an indication of heterogeneity within the types. Some deep valleys of probability can be seen in the central parts of the map in Fig. 4.2, drawing clear boundaries between the types represented in the periphery of the map and those around its center. Some amongst the latter are the ones with a higher level of mixing.

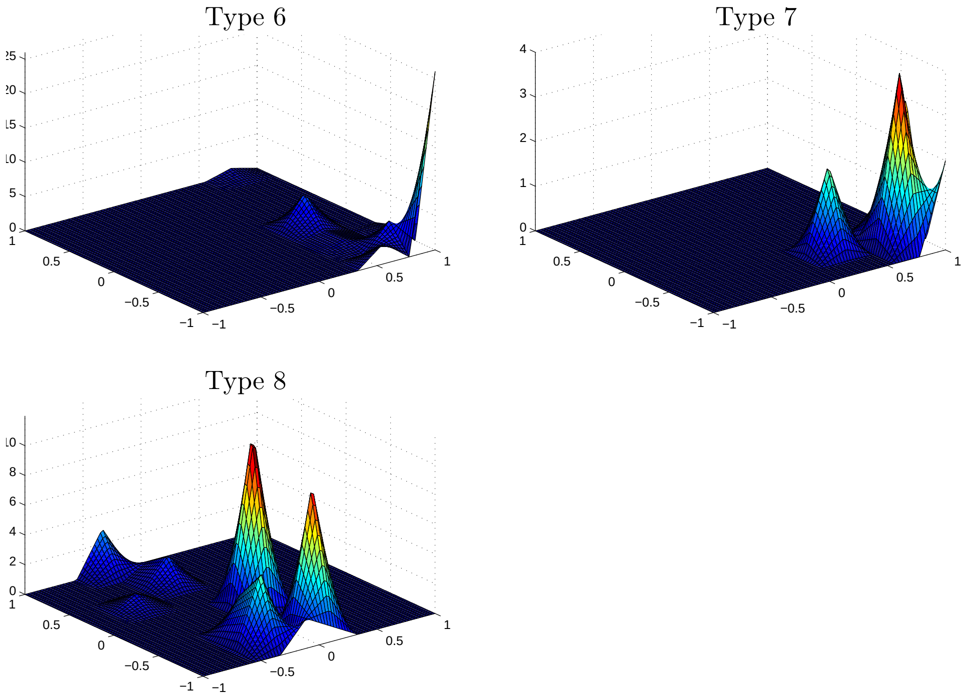
Figure 4.3: CR_c representation maps for all GPCR family C types. Labels: 1: Metabotropic glutamate, 2: Calcium sensing, 4: GABA-B, 5: Vomeronasal, 6: Pheromone, 7: Odorant, 8: Taste. Type 1 (Metabotropic glutamate), the most populated, is well-defined on the top-right corner of the map; type 4 (GABA-B), also isolated and unmixed in the left hand-side of the map; type 6 (Pheromone), strongly focused on the bottom right corner of the map, but partially overlapping with right: type 7 (Odorant). The layout corresponds to that of figure 4.1, although with its viewpoint slightly displaced to the left, to provide some perspective.


Type 1

Type 2

Type 4

Type 5



Our results are consistent with early classification studies using other techniques such as Hidden Markov Models [164], thereby validating the present methodology. Importantly, the proposed method reveals mixing between some receptor subtypes, suggesting its possible applicability to the study of heterodimerization between receptors. Receptor heterodimerization has been confirmed experimentally for a number of receptors [12]. KGTM is shown to help in the exploration of receptors susceptible of heterodimerization and thus be useful in the quest of more potent and safer drugs.

Similarly to CR representation maps, we are also interested in the visualization of the mode projections corresponding to individual subtypes of family C. This representation also lets us explore the same data in different mappings of KGTM:

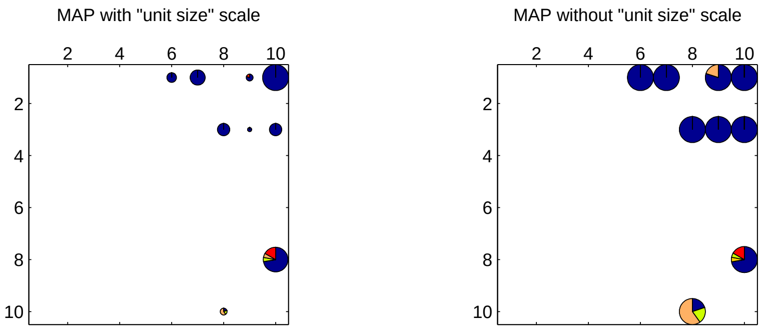


Figure 4.4: Subtype 1 data visualization on a 10 × 10 KGTM representation map, using the mode projection.

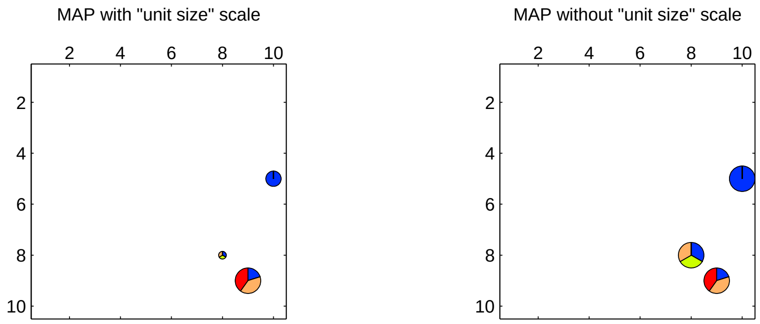


Figure 4.5: Subtype data visualization on a 10 × 10 KGTM representation map, using the mode projection.

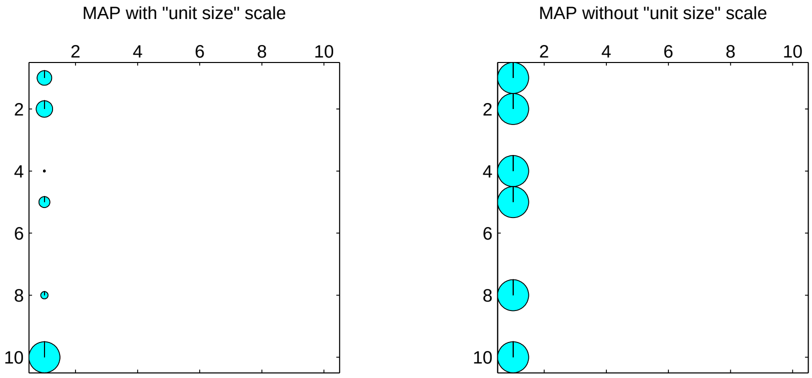


Figure 4.6: Subtype 4 data visualization on a 10 × 10 KGTM representation map, using the mode projection.

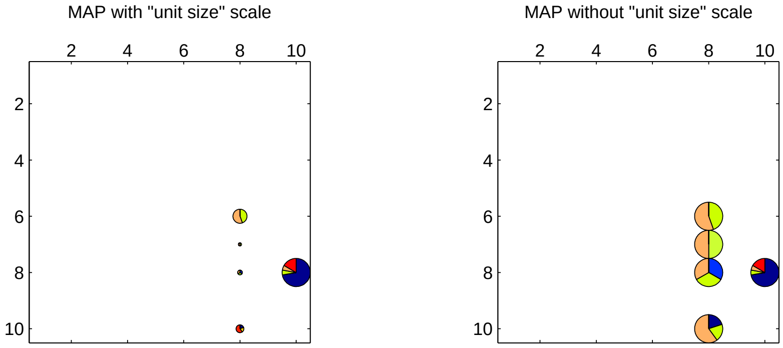


Figure 4.7: Subtype 5 data visualization on a 10 × 10 KGTM representation map, using the mode projection.

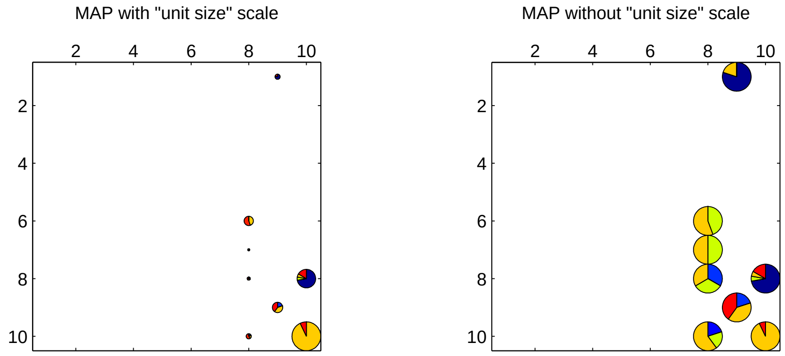


Figure 4.8: Subtype 6 data visualization on a 10 × 10 KGTM representation map, using the mode projection.

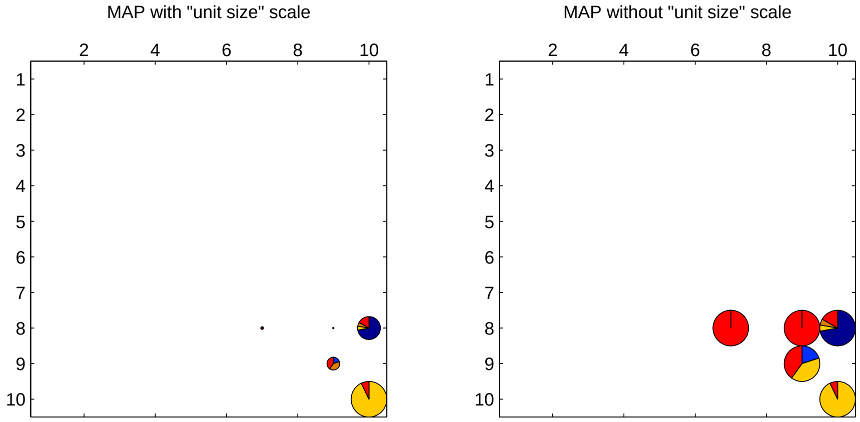


Figure 4.9: Subtype 7 data visualization on a 10 × 10 KGTM representation map, using the mode projection.

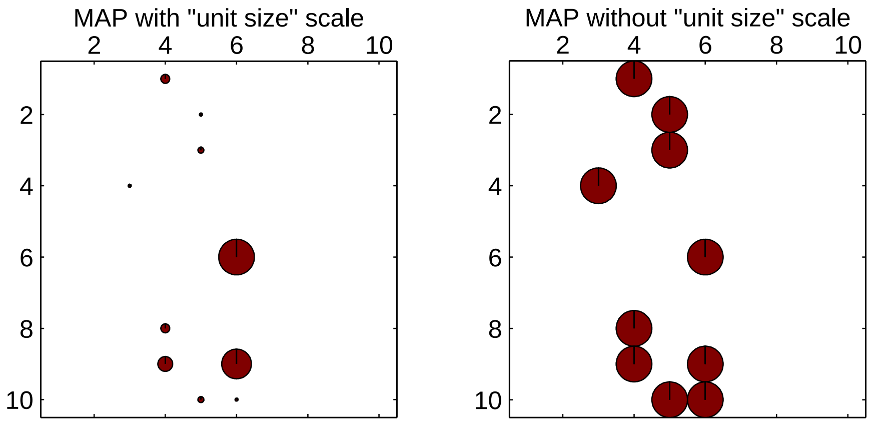


Figure 4.10: Subtype 8 data visualization on a 10 × 10 KGTM representation map, using the mode projection.

4.1.2. Visualization of data set 1 using a standard PT

Before applying the method to construct the PT, data processing was carried out in order to verify the correct location of the clusters once the tree had been created. With that purpose, sequences were labelled as well as the final cluster disposition in KGTM, adding the number of the cluster at the end of the sequence.

The distance method applied to the referred distance matrix BLOSUM62 was the Unweighted Pair-Group Method with Arithmetic Mean (UPGMA)

[183], which examines the structure present in a pairwise distance matrix (or a similarity matrix) and then builds the PT. UPGMA works by progressively clustering the most similar sequences until all the sequences form a rooted tree.

Ultimately, UPGMA yields a distance-based sequence clustering solution in the same sense that KGTM provides one. There are radical differences between them, though. UPGMA is strictly hierarchical in nature and proceeds agglomeratively. It means that once agglomerated, clusters cannot be partitioned any longer throughout the procedure. This introduces a directional bias in the solution. Also importantly, cluster assignments at each level of the tree hierarchy are completely symmetrical; that is, the relative position of a sequence within each cluster is arbitrary, which makes the direct interpretation of proximity not too straightforward, specially for big trees.

On the other hand, KGTM is not hierarchical or agglomerative in nature, which avoids any directional bias. Also, its visualization map makes the assessment of proximity far more intuitive and devoid of any symmetry-related artifacts.

In the following figures, we display the KGTM visualization of each of the GPCR subtypes together with the portion of the PT they correspond to. A visual comparison of both reveals striking similarities.

GPCR subtype 4 (GABA-B) is neatly separated from the rest of types in the KGTM representation (See figure 4.11). The PT reproduces this isolation not only globally (all subtype 4 sequences occupy contiguous tree branch locations) but even to the detail of individual KGTM clusters (each of the 6 clusters allocated by KGTM correspond, quite accurately, to contiguous subregions of the tree).

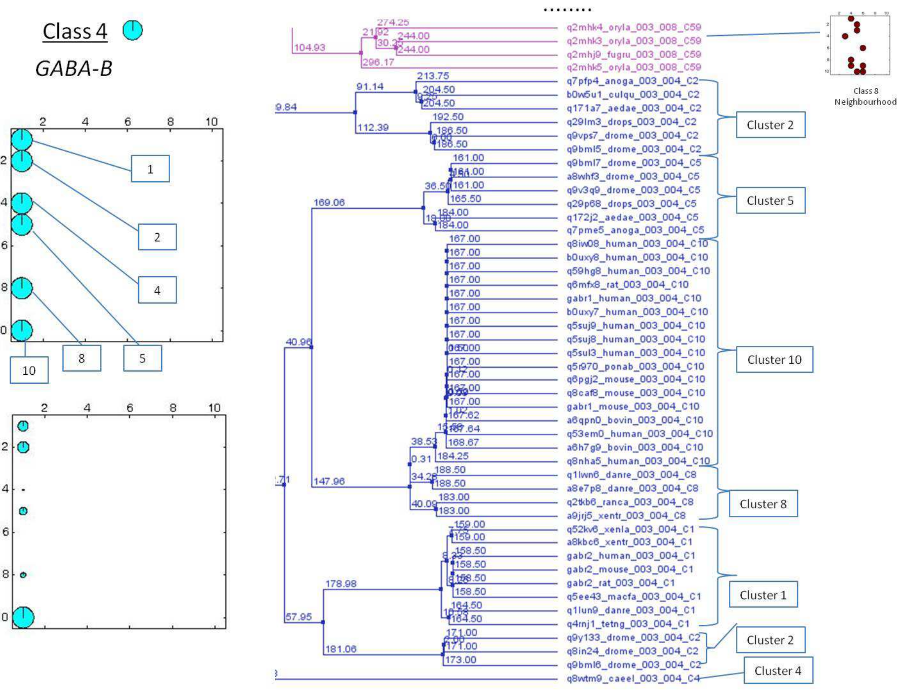


Figure 4.11: Data visualization of subtype 4 (GABA-B), using the mode projection of Eq. 4.1 (left; top: Pie charts with size proportional to the ratio of sequences assigned to them; bottom: without that proportionality); right: its corresponding PT.

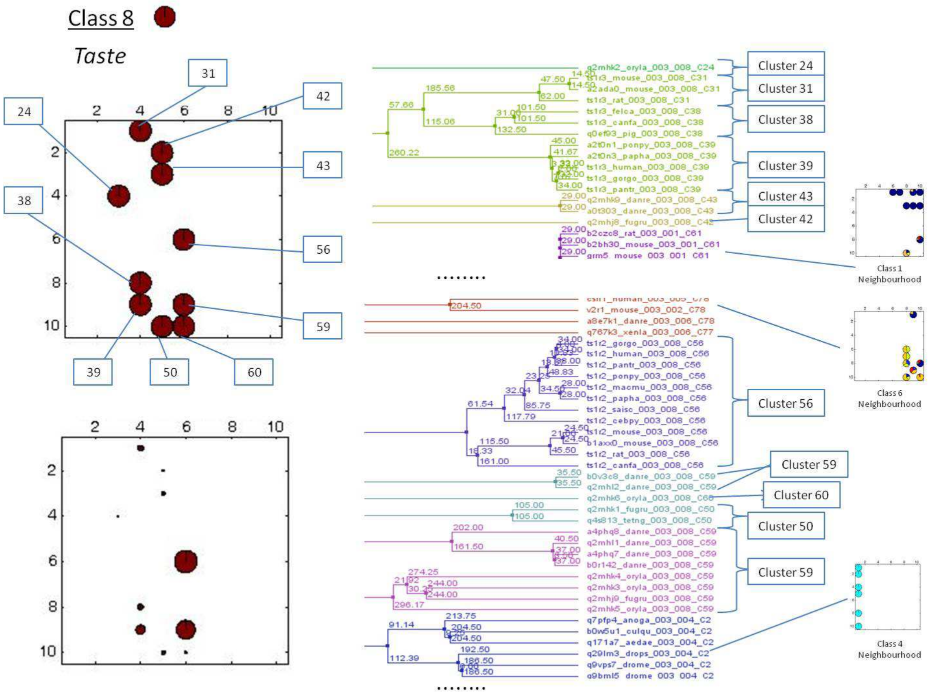


Figure 4.12: Data visualization of subtype 8 (Taste), as in previous figure.

Subtype 8 is also clearly isolated from the rest in the KGTM map, with no mixing in its composition. However, the PT separates it in two clearly differentiated branches. This separation corresponds to two clear cluster locations: one group of clusters located at the top of the KGTM map and the other at the bottom (See Fig. 4.12).

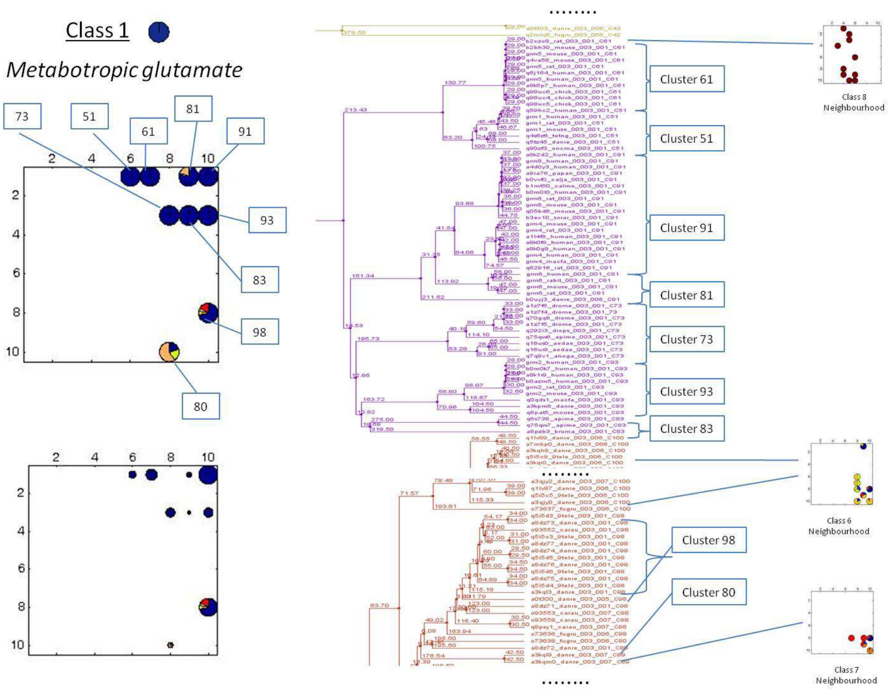


Figure 4.13: Data visualization of subtype 1 (Metabotropic glutamate), as in previous figure.

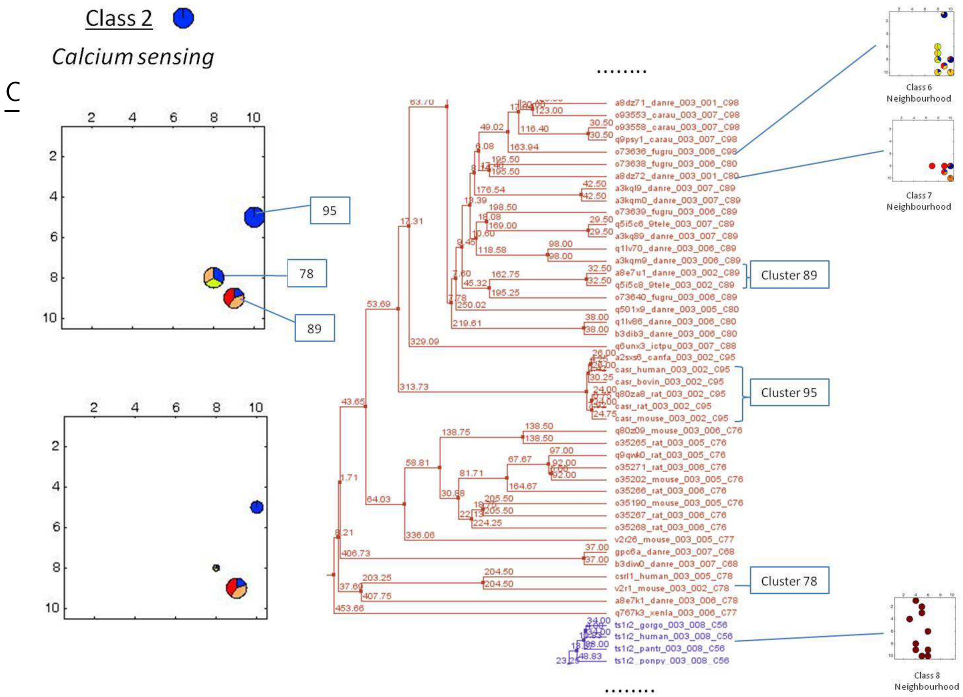


Figure 4.14: Data visualization of subtype 2 (Calcium sensing), as in previous figure.

A clear neighbourhood relationship between class C GPCR subtype 8 and subtypes 4, 1 and 6 is also revealed in both KGTM and the PT. A single sequence belonging to subtype 8 provides us with a very illustrative example: the PT locates it in a very differentiated tree branch, at the top of the tree in Fig. 4.12. By itself, it forms KGTM cluster 24, which is clearly isolated from the rest of subtype 8.

The GPCR subtype 1 also has a very compact phylogenetic representation that matches overall with the grouping provided by the KGTM model (See figure 4.13). In particular, we find some isolated subtype 1 sequences in the PT, located between subtype 6 and subtype 7 sequences, which are assigned to the isolated location of clusters 80 and 98 in the KGTM map.

Finally, we have subtype 2,5,6 and 7 which show a far more heterogeneous structure both in the PT and in the KGTM map, although they still preserve neighbouring relations in both representations (See figures 4.14,4.15,4.16 and

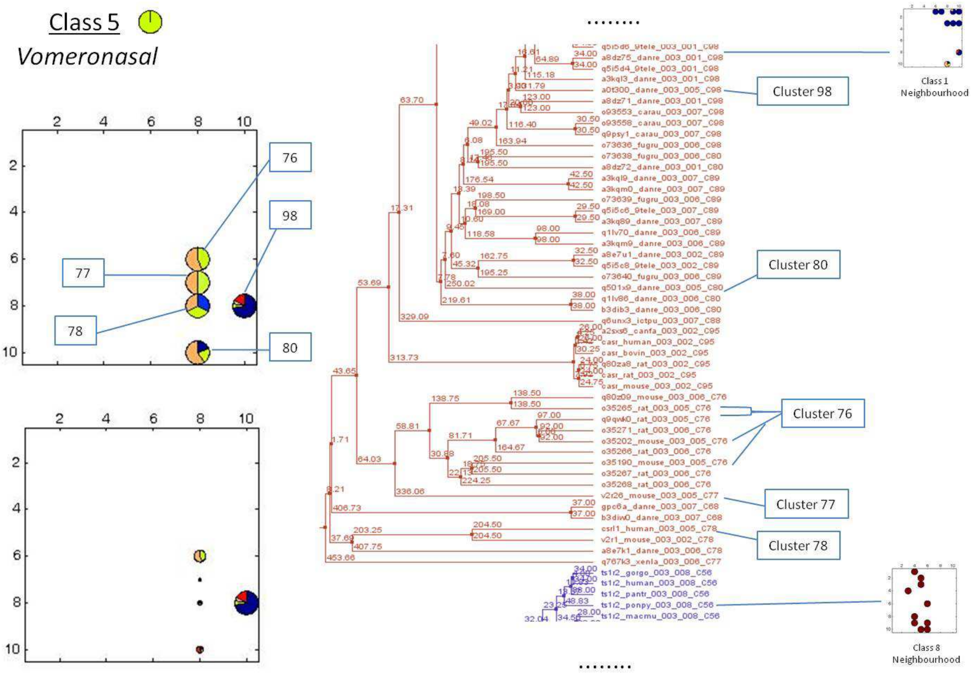
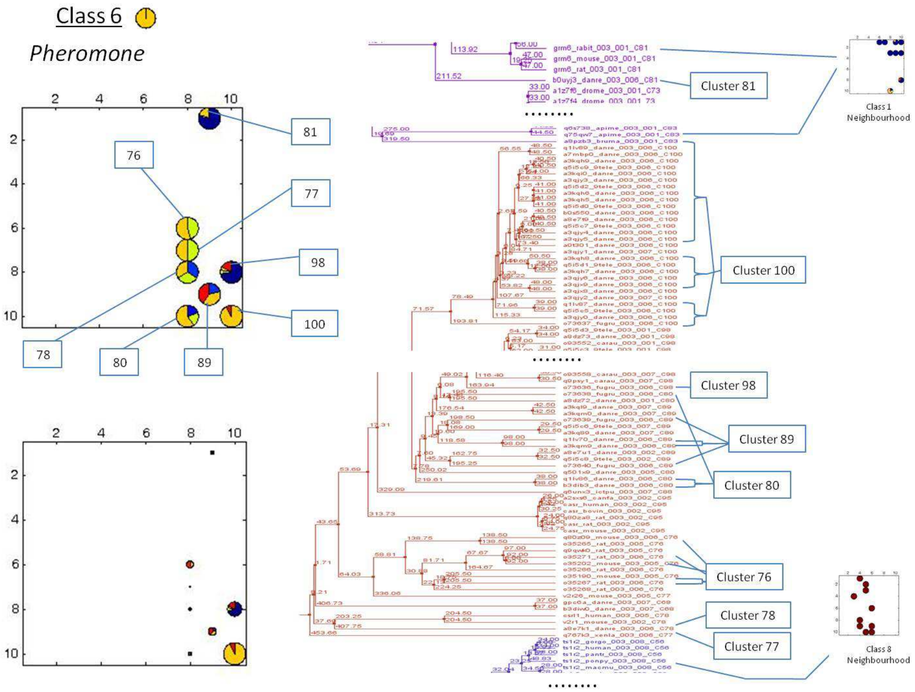


Figure 4.15: Data visualization of subtype 5 (Vomeronasal), as in previous figure.

4.17).

Figure 4.18 shows the complete PT representation of the class C GPCR data analysed in this thesis. The colors in the tree are automatically generated by the software. Same color is assigned to close leaves (sequences) and branches (groups of sequences) of the tree, according to the evolutive distance between sequences. These distances are the numbers attached to the branches. The software also automatically plots a red line which establishes the depth from which the color grouping starts. Individual sequences in the leaves of the tree are labelled according to three items: their ID , the family and the type (e.g.,



Fig

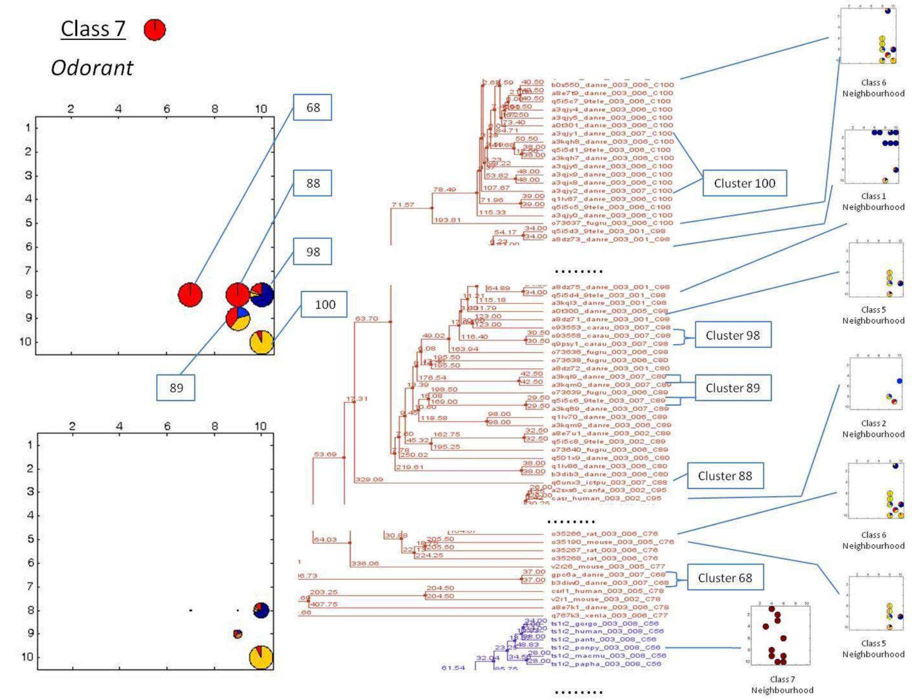
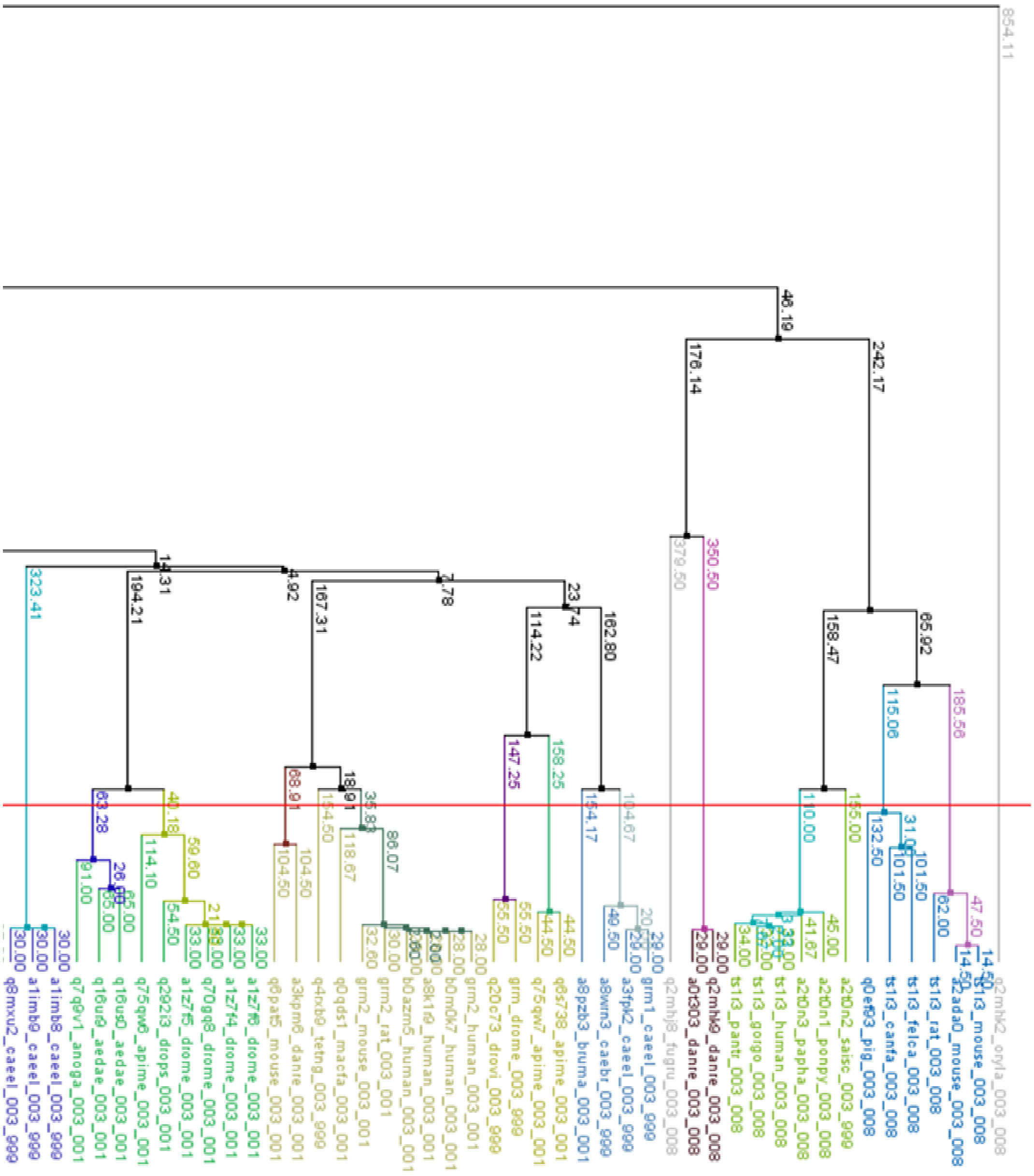


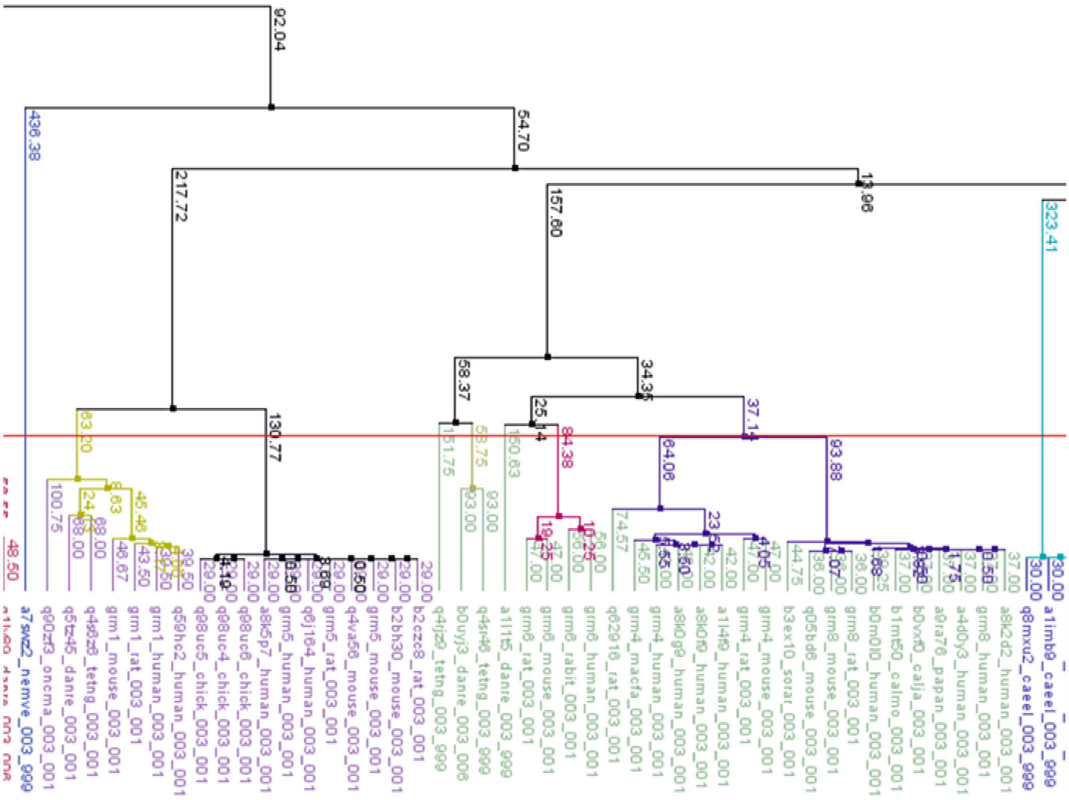
Figure 4.17: Data visualization of subtype 7 (Odorant), as in previous figure.

sequence $ts1r3\_mouse\_003\_001$ indicates ID: $ts1r3\_mouse$; family: 003 (C); and subtype: 001).

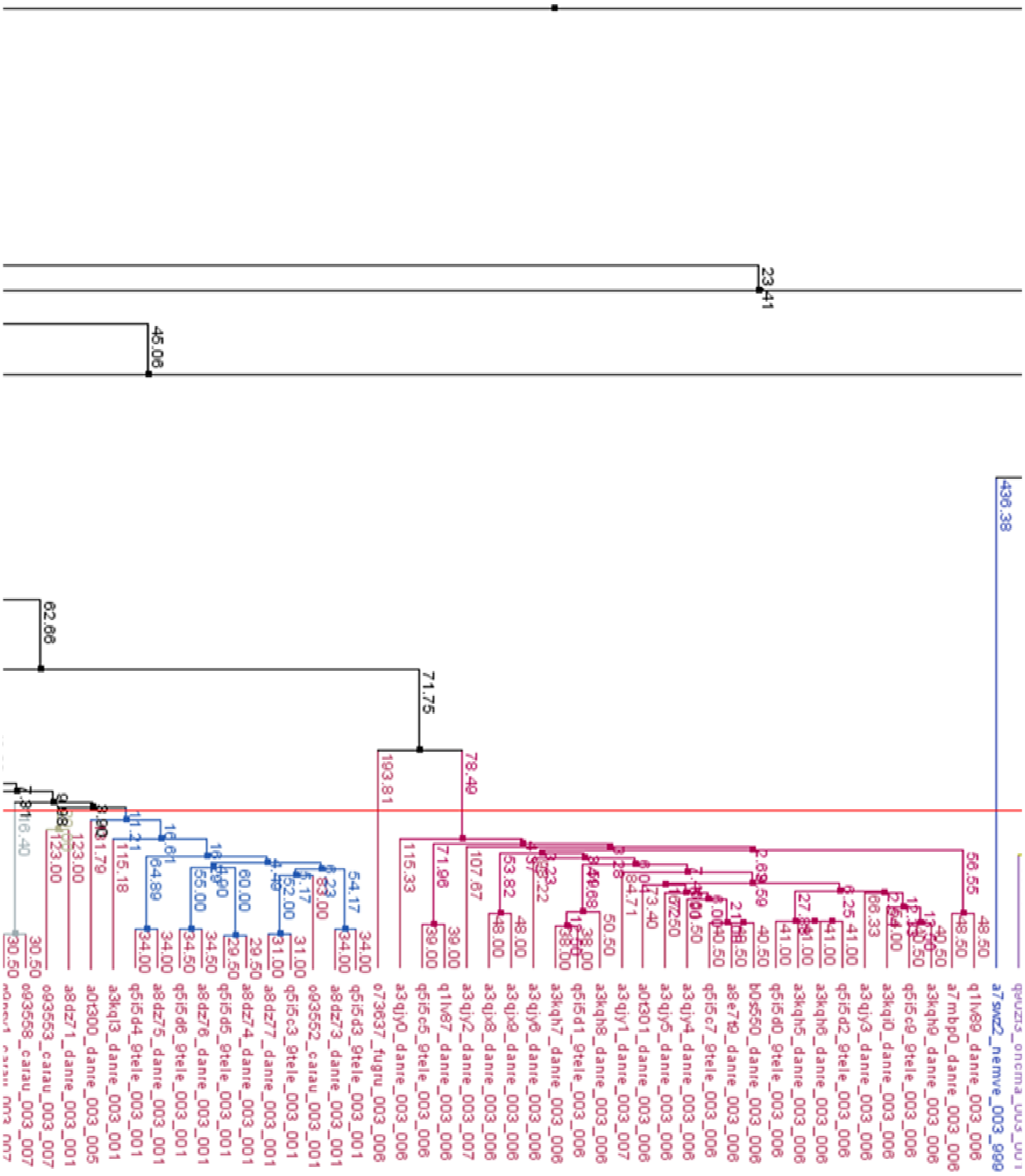
Figure 4.18: The figure has been split due to space limitations. - continues on the next page -



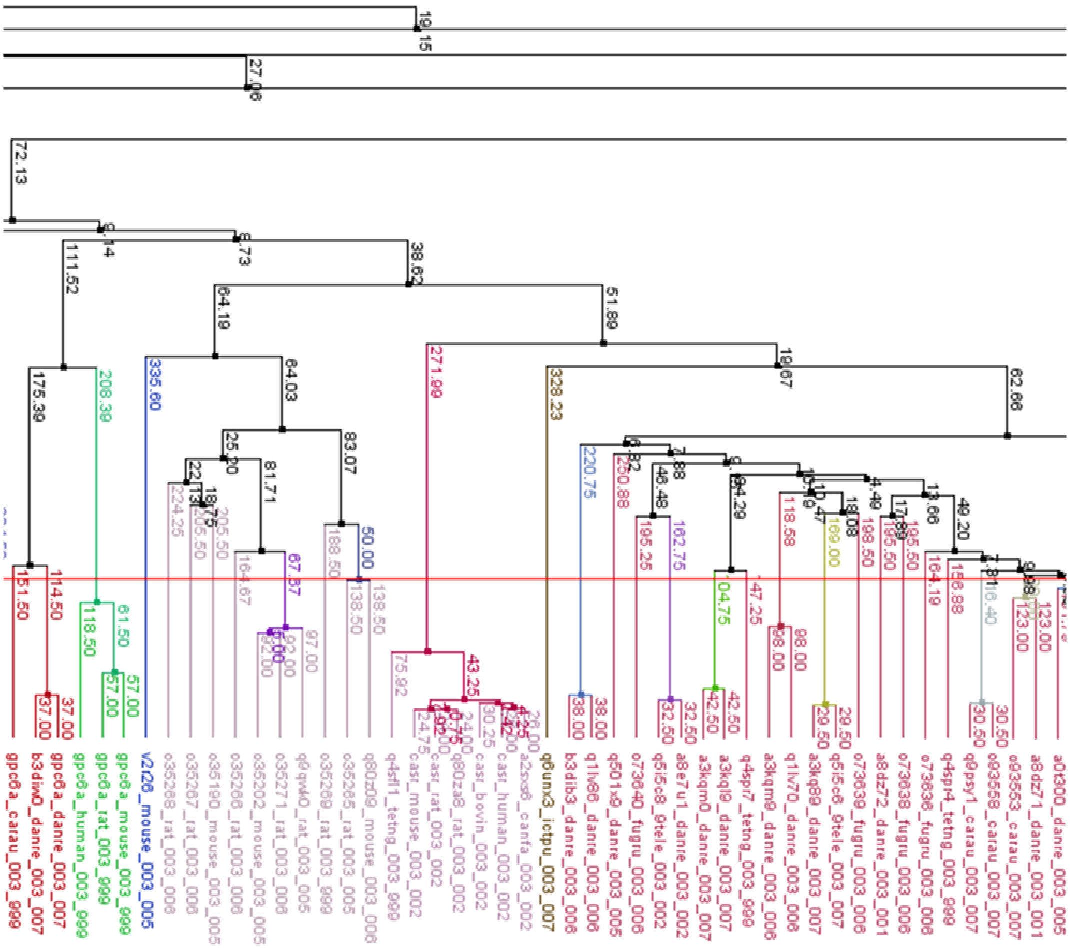
- continues on the next page -



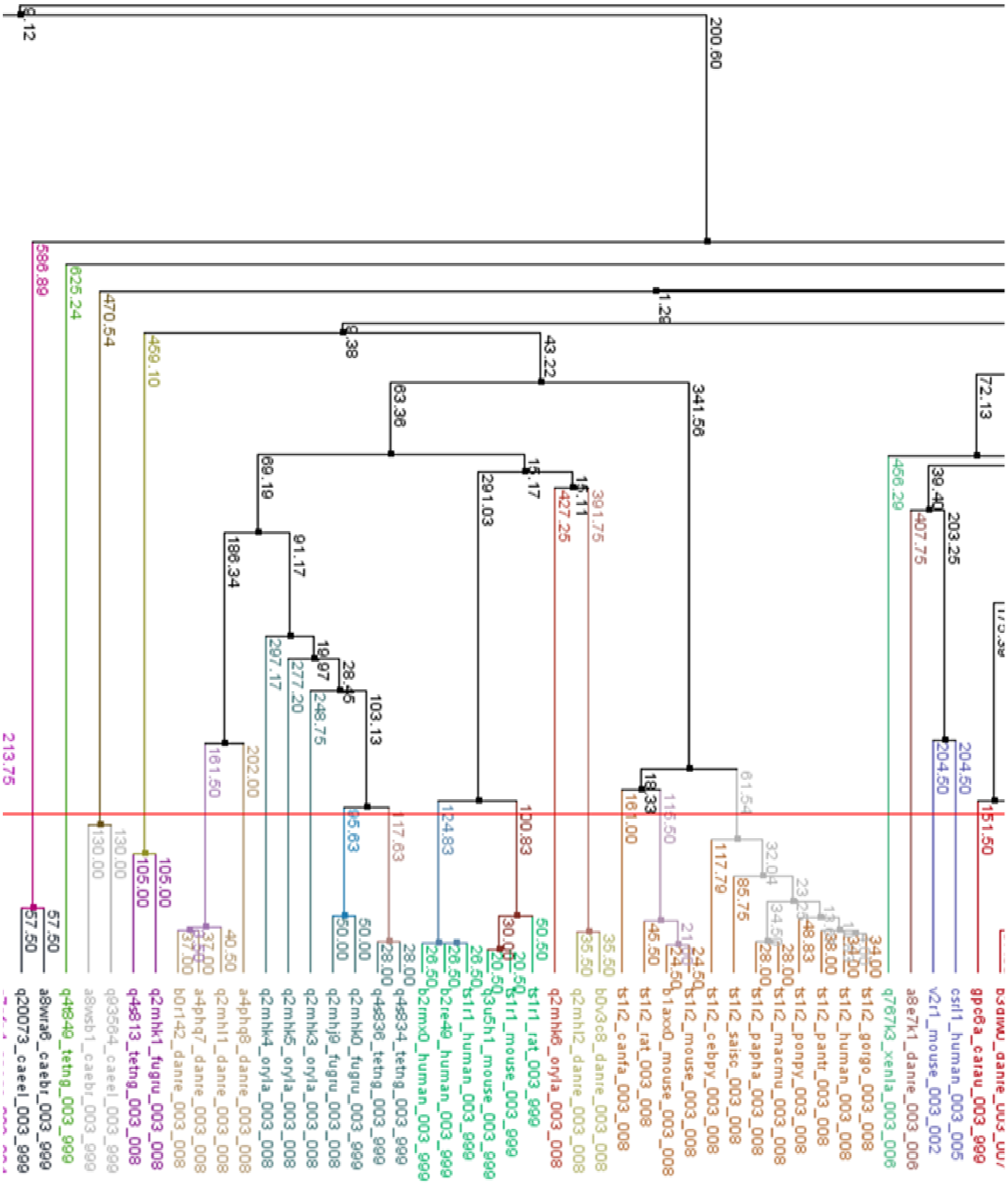
- continues on the next page -



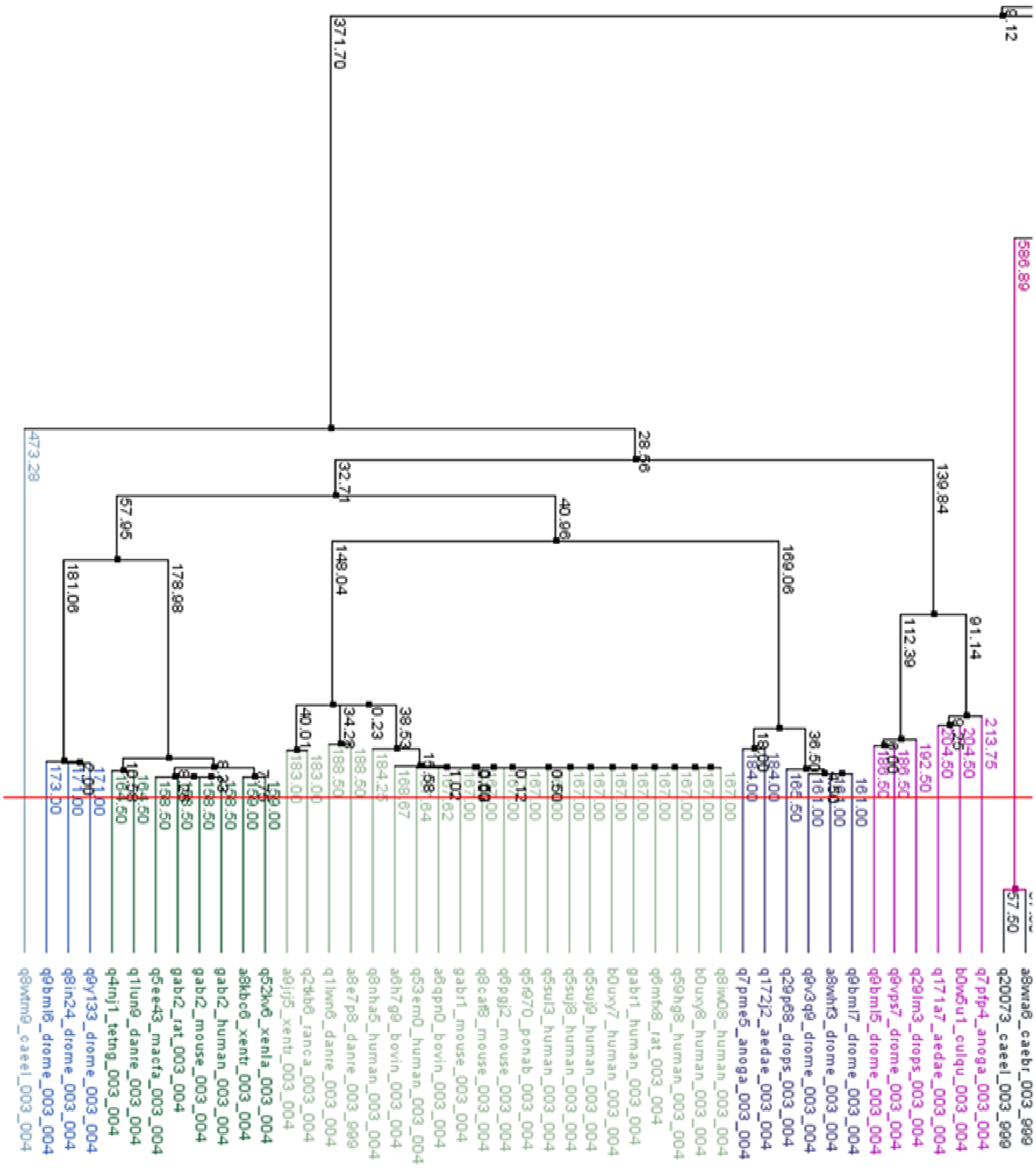
- continues on the next page -



- continues on the next page -



- continues on the next page -



4.2. Measures for Quantitative Assessment of the Data Grouping Procedure

In order to complement the so far mostly qualitative exploratory visualization of the class C GPCR sequences, we describe here several measures for a quantitative assessment of subtype overlapping: Entropy, Distribution Consistency (DC) and Distance Consistency (DSC).

Entropy: In our experiments, entropy should be understood as a measure of class C heterogeneity. When (K)GTM map areas are completely subtype-specific (that is, when no two sequences of different subtypes are assigned to the same (K)GTM latent point), the corresponding entropy will be zero, whereas high entropies will characterize highly overlapping subtypes (with sequences of different subtypes very mixed in the same (K)GTM latent point).

Generally speaking, entropy depends on the probability that the model attributes to the source. In the case of (K)GTM, the total entropy for a given latent point k in the visualization space will be expressed as

$$S_k = -\sum_{j=1}^{C} p_{kj} ln p_{kj}, \tag{4.2}$$

where j is one of the seven GPCR Class C subtypes and $p_{kj} = \frac{m_{kj}}{m_k}$, where m_k is the number of sequences assigned to cluster k and m_{kj} is the total number of sequences assigned to cluster k that belong to subtype j.

Then, the **Weighted-Average Entropy** assigned to each subtype j, for all units in the visualization map could similarly be defined as

$$E^j_{wa} = -\sum_{k/m_{kj}>0} S_k \frac{m_k}{N_j}, \tag{4.3}$$

where N_j is the number of sequences from a subtype j. Finally, the total Weighted-Average Entropy is defined as

$$E_{wa} = \sum_k S_k \frac{m_k}{N}, \tag{4.4}$$

where N is the total number of sequences in the data set under analysis.

Distribution Consistency (DC): The DC measure [10] quantitatively reproduces a visual assessment method based on a weighted average of the latent point entropies and is normalized in order to give a score between 0 and 100, where the highest score means the higher separation. It is definition contains the E_{wa}:

$$DC = 100 - 100\frac{E_{wa}}{log(C)}, \tag{4.5}$$

where C is the number of classes in the dataset. Note that this is an overall discrimination measure (and therefore does not provide individual measures per subtype) to be compared to E_{wa}. Unlike E_{wa}, the higher the result, the better the discrimination capabilities it reveals.

Distance Consistency (DSC): The DSC measure [180] is the proportion of data points x_n whose nearest class-center-of-mass belongs to the same class as x_n.

Let us define the centroid distance (CD). It describes the property of class members that the distance $d(x, centr(C_i))$ to its class centroid should be always

minimal in comparison to the distance to all other centroids.
Then, DSC is defined as:

$$DSC = 100 - 100\frac{MCNC}{N}, \tag{4.6}$$

where $MCNC = |x'_n \in \Re^2 : CD(x'_n, centr(C(x'_n)) \neq true|$ are the misclassified examples of the nearest centroids and $centr(C(x'_n)) = \frac{1}{|C(x'_n)|}\sum_{x' \in C(x'_n)} x'$, where $x'_n = \arg\max_{\{k_n\}} r_{kn}$ are the projections of x_n in latent space mode projection.

Again, this is an overall discrimination measure to be compared to E_{wa}. Like DC, the higher the result, the better the discrimination capabilities it reveals.

4.2.1. Results and discussion

In the next section, are reported the results of a quantitative estimation of subtype overlapping using the visual discrimination assessment measures previously defined. This estimation allows us to focus the analysis not on subtype discriminability but, instead, on subtype overlapping and its hypothetical consistency over different sequence transformations. We go beyond the previous experiments reported for data set 1, as we compare the visualizations provided by GTM and KGTM.

Visualization of data set 2 with GTM and KGTM

The standard GTM is used here to model and visualize the AAC-transformed *unaligned* sequences, while KGTM is used here to model and visualize the MSA-transformed sequences (this is, thus, an *unaligned* vs. *aligned* sequences experimental setting).

Figure 4.19 visualizes the class C GPCR AAC-transformed data set using the *posterior mode projection* representation for the standard GTM in a 15×15 ($K = 225$) latent grid. Figure 4.19 (left) reveals the heterogeneity of the GPCR groupings, with some areas of the mapping concentrating most sequences, whereas Figure 4.19 (right), with the relative map unit-size effect removed, suggests that some subtypes are more clearly mixed than others.

Figure 4.20 similarly visualizes the MSA-transformed data set using KGTM. This time, the level of overlapping seemingly diminishes and subtypes appear more clearly separated. This is neatly reflected by the mapping of the cumulative *responsibilities* $CR_k = \sum_{n=1}^{N} r_{kn}$ in Figure 4.21, where the probabilities of data assignment are concentrated in limited spaces of the nonlinear mapping that correspond to the biggest clusters in Figure 4.19 (left).

Quantifying data set 2 overlapping with Visual Discrimination Assessment Measures

For the exploratory visualization of the Class C GPCR sequences, is applied here an entropy-based measure that is suitable for discrete clustering visualizations such as those provided by the GTM variants. The entropy is a measure of Class C heterogeneity. Table 4.3 summarizes the entropies per subfamily E_j and the E_{wa} for each of the transformed data sets in our study.

The entropy results reported in Table 4.1 only partially corroborate our starting hypothesis. The overall entropy of the KGTM representation of the MSA-transformed sequences is lower than the corresponding entropy of the GTM representation of the AAC transformation, both for the complete sequences and for the N-terminus. In all cases, the complete sequences yield lower entropies

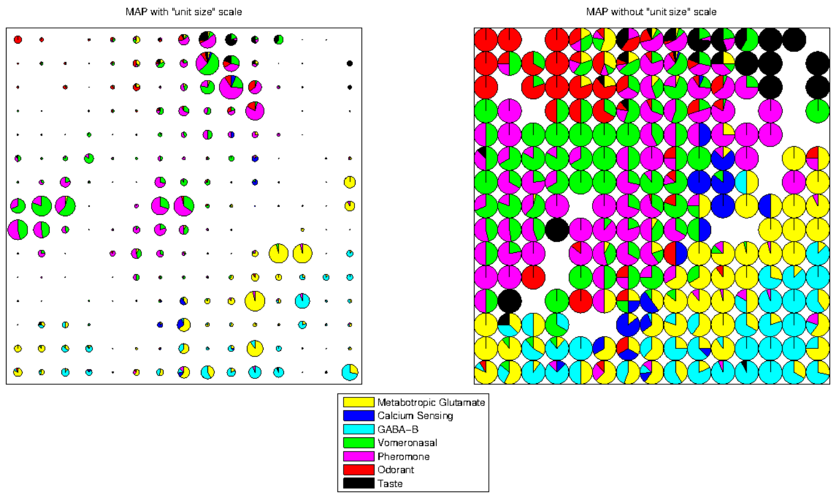


Figure 4.19: Visual map of the standard GTM-based *posterior mode projection* of the GPCR unaligned data, transformed using AAC. Each pie chart is a partition by subtype (colour-coded as described in the legend) of the GPCR sequences mapped onto a given latent space point u_k of the 15×15 grid. Left) Size of pie charts is scaled in proportion to the ratio of sequences mapped onto them. Right) the same map without scale to visualize emphasize the partition of small pie charts.

than the N-terminus; this means that the N-terminus only partially retains the subtype discrimination capabilities of the complete sequence. The inspection of the entropies per subtype reveals a less clear-cut picture. For the GTM with AAC, the use of the N-terminus increases the entropy for the easier-to-discriminate subtypes (mGlu, CS, GB, Ta), while it decreases the entropy for the most overlapping ones (VN, Ph, Od). For the KGTM with MSA is precisely the other way around: the use of the N-terminus decreases the entropy for the easier-to-discriminate subtypes and increases the entropy for the most overlapping ones. In any case, MSA keeps the entropies of the overlapping subtypes at rather low values.

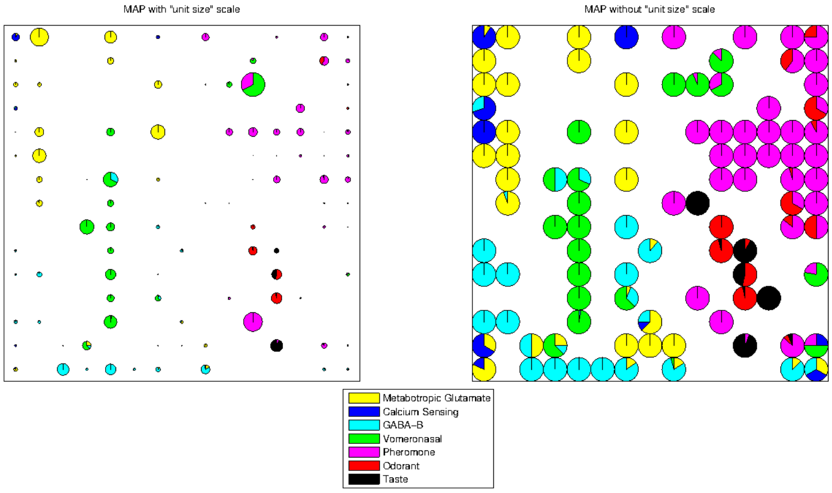


Figure 4.20: Visual map of the KGTM-based *posterior mode projection* of the Class C GPCR data, transformed using MSA. Representation as in Fig.4.19

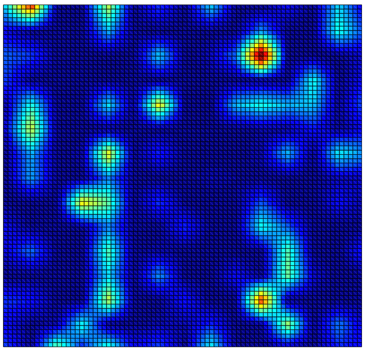


Figure 4.21: Colour-coded visual representation of the cumulative *responsibility* for KGTM. Dark red areas correspond to the highest probability of sequence assignment and, therefore, to dense concentrations of GPCR sequences, whereas deep blue areas correspond to the lowest probability (data empty spaces).

Visualization of data sets 2 and 5 with GTM and KGTM

As data set 5 is a subset of data set 2, we hypothesize that the results of class C GPCR subtype discrimination should differ depending on whether we

Table 4.1: Entropies for each of the 7 subtypes for N-terminal Domain and Complete GPCR.

GPCR Complete								
	MGlu	**CS**	**GB**	**VN**	**Ph**	**Od**	**Ta**	E_{wa}
AAC	3.65	0	2.75	15.05	15.90	4.36	0.43	0.37
MSA	3.43	1.89	3.56	2.34	3.63	2.99	0.91	0.18
N-terminal Domain								
	MGlu	**CS**	**GB**	**VN**	**Ph**	**Od**	**Ta**	E_{wa}
AAC	5.91	0.35	3.40	9.52	12.23	4.20	2.38	0.46
MSA	2.98	1.42	1.42	8.10	9.07	4.7	0	0.26

use the complete primary sequence or, instead, we use only the extracellular N-terminus domain of the receptor. Related to this, we also hypothesize that the N-terminus should be almost as good as the complete sequence in terms of subtype discrimination. The reason for this lies on VFT including the site where endogenous ligands for class C GPCRs bind and, as a consequence, a diversity in AA sequence is expected. A secondary hypothesis is that these differences should intuitively be observed through manifold learning-based visualization.

Our experiments are organized according to two different dimensions. First, we analyzed the available sequences according to two different transformations using two different methods: unaligned sequences are transformed according to the AAC method and analyzed using the standard GTM, while KGTM is used to analyze sequences transformed by MSA. Second, we use two approaches to assess the results: exploratory visualization for a qualitative interpretation of the global (sub)structure of subtypes, complemented by a quantitative assessment of the level of subtype discrimination, based on an entropy measure.

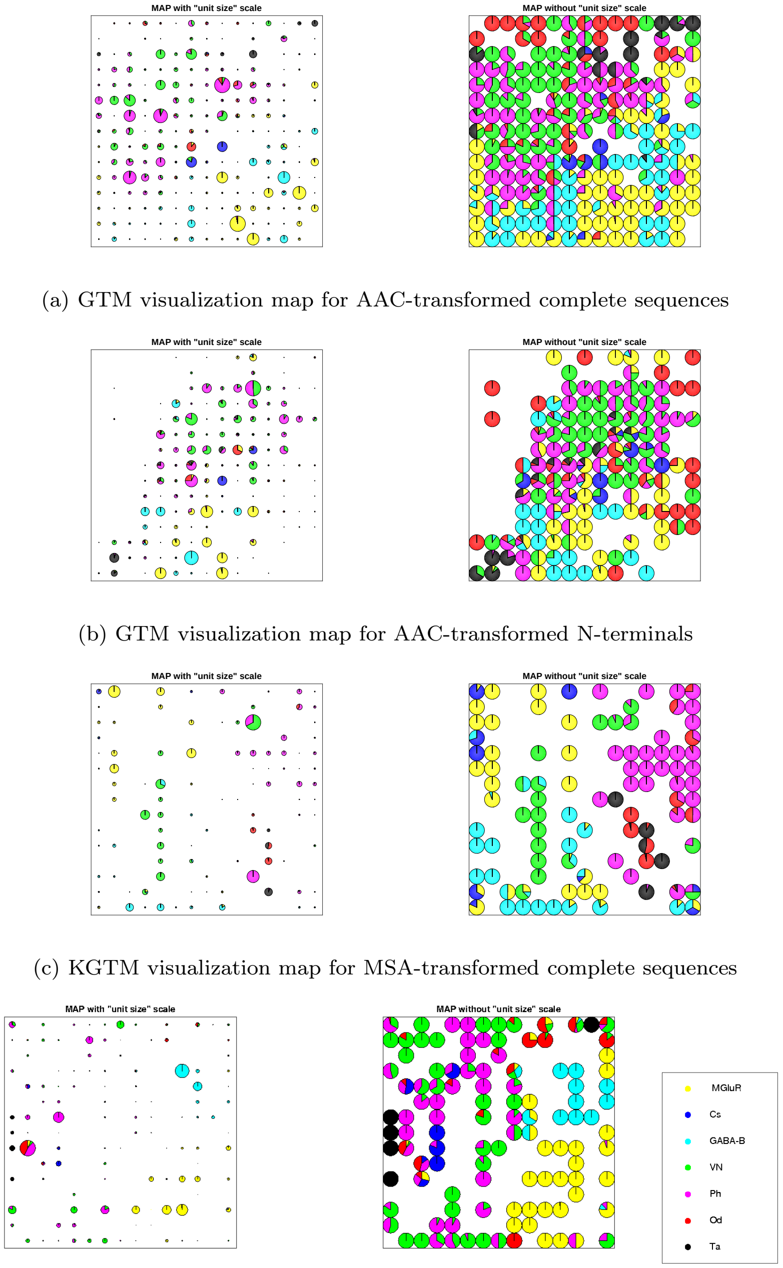
The mode projection of the AAC-transformed data on the standard GTM visualization map is shown in Fig.4.22a for the complete Class C GPCR sequences and in Fig.4.22b for the extra-cellular N-terminus of the same sequences.

Correspondingly, the mode projection of the MSA-transformed data on the KGTM visualization map is shown in Fig.4.22c for the complete class C GPCR sequences and in Fig.4.22d for the extra-cellular N-terminus of the same sequences.

In order to visually assess the level of subtype mixing in each of (K)GTM latent points, modes are again represented as pie charts. These mode projections are, in the end, a simplified representation in which each sequence is mapped to the latent point of highest responsibility r_{kn}. We also again use the richer probabilistic information provided by the model to inspect the *responsibility maps* of individual sequences through visualization of the distribution of r_{kn} values on the (K)GTM maps. Some examples are shown in Fig. 4.24.

The mode projections for all data sets in Figure 4.22 reveal some striking differences. Overall, the GTM representation of the AAC-transformed sequence projections is far more distributed than that of the KGTM of the MSA-transformed sequence projections, with many latent points taking responsibility for only a few sequences. Interestingly, and specially for the GTM, the N-terminus projection is much more compact than that of the complete sequence, involving far fewer latent points. In all cases, a limited number of latent points concentrates a relatively large number of sequences; this is particularly the case in the KGTM MSA-transformed representation.

The examples of individual r_{kn} in Figure 4.24 correspond to cases from different sub-families in which the probability of assignment of sequences to latent points is clearly multi-modal. This illustrates the way the models handle uncertainty. Multi-modal cases with lower maxima are most frequent in subtypes with high levels of overlapping, such as VN, Ph and Od.



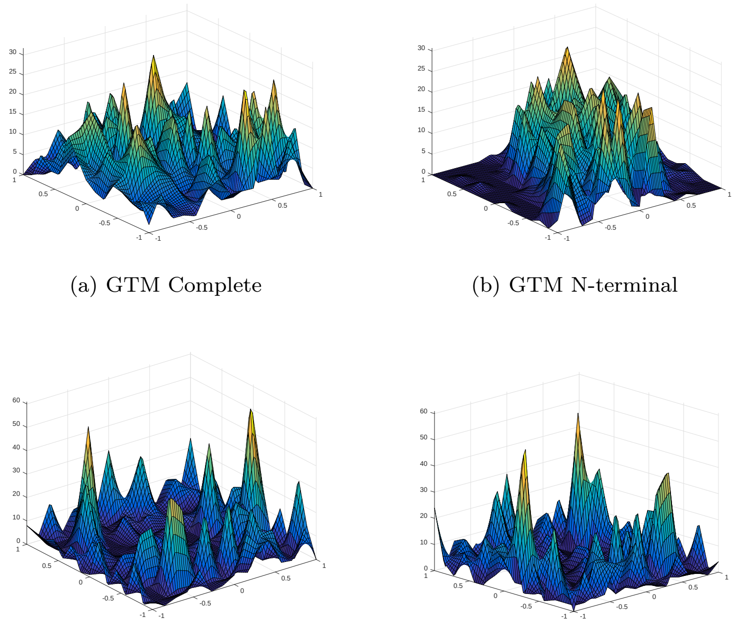
(a) GTM visualization map for AAC-transformed complete sequences

(b) GTM visualization map for AAC-transformed N-terminals

(c) KGTM visualization map for MSA-transformed complete sequences

(d) KGTM visualization map for MSA-transformed N-terminals (left and centre).

Figure 4.22: Visualization maps of the different data *mode projections*. The left and right columns display the same data representation; their difference is that, in the maps on the left, the size of the pie chart encodes the ratio of sequences assigned to a given latent point, therefore providing visual clues about the spatial distribution of relative data density. Subfamily labels for all maps are shown in the bottom-right legend.



(a) GTM Complete

(b) GTM N-terminal

(c) KGTM Complete

(d) KGTM N-terminal

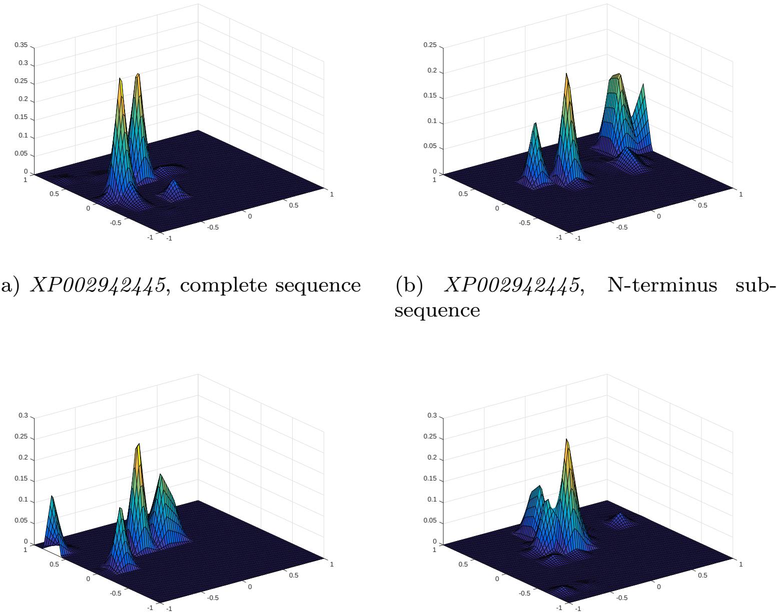
Figure 4.23: Map of the CR (vertical axis) over the (K)GTM latent visualization space for all datasets.

Table 4.2: The E_{wa}, DC and DSC overall measures for the complete GPCR, the N-terminal domain and the 7TM domain.

	GPCR Complete			N-terminal Domain			7TM Domain		
	E_{wa}	**DC**	**DSC**	E_{wa}	**DC**	**DSC**	E_{wa}	**DC**	**DSC**
AAC	0.37	80.86	39.21	0.46	76.17	36.98	0.58	70.41	28.99
MSA	0.18	90.54	52.32	0.26	86.65	48.40	-	-	-

Quantifying data set 2 overlapping with Visual Discrimination Assessment Measures

Table 4.2 summarizes the overall E_{wa}, the DC and DSC measures for the complete GPCR, the N-terminal extra-cellular domain and the 7TM domain, for some of the data transformations and GTM variants.



(a) *XP002942445*, complete sequence

(b) *XP002942445*, N-terminus subsequence

(c) *XP002940939*, complete sequence

(d) *XP002940939*, N-terminus subsequence

Figure 4.24: Visualization of the GTM responsibility r_{kn} for some example AAC-transformed sequences (standard database names included) from subfamilies VN (*a* and *b*) and Ph (*c* and *d*).

In more detail, Table 4.3 summarizes the entropies E_j, the DC and the DSC per subtype, for some of the transformed datasets in our study. Statistical tests to assess the significance of the differences between results are reported in Table 4.4. They include tests comparing results obtained with KGTM and MSA against those obtained with GTM and AAC, and also comparing results obtained using the complete sequence and specific domains. T-tests were used for E_{wa} and DC, while Fisher's test was used for DSC.

The results reported in Table 4.2 indicate that the use of KGTM with the MSA transformation provides better discrimination results than the use of GTM with the AAC sequence transformation, in terms of both the E_{wa} and DC

Table 4.3: Subtype entropies, DC and DSC measures for the complete GPCR, the N-terminal domain and the 7TM domain.

GPCR Complete - AAC							
	MGlu	**CS**	**GB**	**VN**	**Ph**	**Od**	**Ta**
E^j_{wa}	0.36	0.59	0.42	0.56	0.58	0.67	0.51
DC	81.31	69.88	78.45	71.28	70.41	65.56	73.70
DSC	44.73	85.41	31.25	36.91	33.67	15.68	83.07
GPCR Complete - MSA							
	MGlu	**CS**	**GB**	**VN**	**Ph**	**Od**	**Ta**
E^j_{wa}	0.21	0.51	0.34	0.33	0.21	0.37	0.31
DC	89.36	73.84	82.48	82.84	89.07	81.10	84.04
DSC	26.49	79.16	67.78	36.62	76.78	39.21	78.46
GPCR NT Domain - AAC							
	MGlu	**CS**	**GB**	**VN**	**Ph**	**Od**	**Ta**
E^j_{wa}	0.53	0.74	0.35	0.63	0.65	0.84	0.89
DC	72.80	62.11	81.94	67.53	66.43	56.63	54.02
DSC	51.42	48.89	25	45.73	17.41	30	65.07
GPCR NT Domain - MSA							
	MGlu	**CS**	**GB**	**VN**	**Ph**	**Od**	**Ta**
E^j_{wa}	0.36	0.64	0.24	0.46	0.47	0.87	0
DC	81.27	67.17	87.73	76.12	75.87	55.01	100
DSC	90.07	75.55	94.87	8.53	23.12	8.75	96.83
GPCR 7TM Domain - AAC							
	MGlu	**CS**	**GB**	**VN**	**Ph**	**Od**	**Ta**
E^j_{wa}	0.65	0.76	0.73	0.79	0.81	0.89	0.84
DC	66.78	61.18	62.50	59.23	58.17	53.97	56.99
DSC	43.62	80	38.46	2.73	27.02	5	66.67

measures. These differences are statistically very significant according to the t-tests compiled in Table 4.4 (first two rows). They are still very significant in terms of Fisher's test with *DSC* for NT and somehow less significant for the complete sequence. The use of KGTM with MSA is therefore validated. Note that this conclusion is consistent with the qualitative visual assessment previously discussed.

Results also provide support to the first of the working hypotheses H2.3 and H2.4 stated at section 1.2.1, which stated that the results of Class C GPCR subtype discrimination, as seen from the natural structure of the primary sequence data transformations revealed by unsupervised DR techniques, should

Table 4.4: Results of statistical tests applied to the comparison of results: both to the comparison of DR methods with data transformation and to the comparison of the use of different domains or the whole sequence. The use of p* indicates that the test of differences is moderately significant. The use of p**, in turn, indicates that the differences are not statistically significant according to the test.

STATISTICAL TESTS			
Comparison **Methods/Domains**	E_{wa}T-Test p-value	**DC T-Test** p-value	**DSC Fisher's -Test** p-value
MSA-Complete vs. AAC-Complete	p < 0.0001	p < 0.0001	p* = 0.0390
MSA-NT vs. AAC-NT	p < 0.0001	p < 0.0001	p < 0.0001
MSA-Complete vs. MSA-NT	p < 0.0001	p < 0.0001	p < 0.0001
AAC-Complete vs. AAC-NT	p < 0.0001	p < 0.0001	p** = 0.2381
AAC-NT vs. AAC-7TM	p < 0.0001	p < 0.0001	p < 0.0001
AAC-Complete vs. AAC-7TM	p < 0.0001	p < 0.0001	p < 0.0001

differ depending on whether we used the complete primary sequence of these membrane proteins or, instead, we used only the extracellular N-Terminus or the 7TM domain.

The results in Table 4.2, corroborated by the t-tests for the E_{wa} and DC measures and Fisher's test for DSC in Table 4.4 (third to fifth rows), indicate significant discrimination capabilities, with the complete sequence providing significantly better discrimination than the N-terminal domain and N-Terminus, in turn, better discrimination than the 7TM domain. This is the case both for KGTM with MSA for all measures and for GTM with AAC for all measures but DSC, for which the difference is not deemed to be significant.

These very same results mostly disqualify hypothesis H2.4, which stated that the use of the N-Terminus on its own, given the particularities of this domain, should yield comparable results to the complete sequence in terms of subtype discrimination. Discrimination appears to be significantly better when using the complete sequence, although this statistical significance is lost according to DSC when using the AAC transformation. Supervised analysis using SVMs in [108] with AAC transformed data also yielded a slight advantage for the use of the whole sequence (94% accuracy) over the use of the N-Terminus (93%).

A more nuanced interpretation can be made when exploring the entropy, DC and DSC values per class C subtype, as reported in Table 4.3. KGTM with MSA yields better results in most subtypes than GTM with AAC for E_{wa} and DC, but results for DSC are mixed: for the complete sequence, KGTM with MSA yields better results for subtypes that are comparatively more difficult to discriminate such as GB, Ph and Od, whereas for the N-Terminus, it yields better results only for the easier to discriminate, such as MGlu, CS and Ta.

A similarly inconsistent pattern occurs for DSC when comparing results obtained using the complete sequence with results obtained using only the N-Terminus, both for KGTM and GTM. This is overall an indication that DSC is not an adequate metric for this particular problem.

The results reported in Table 4.3 are also overall consistent with those obtained in previous studies using supervised methods to analyze the same class C GPCR database, applying SVM with feature selection [107], where the following Matthew's Correlation Coefficient (MCC) values (maximum value 1) were obtained per subtype: MGluR (0.95), CS (0.93), GB (0.98), VN (0.89), Ph (0.86), Od (0.79) and Ta (0.99). This is a clear indication that, beyond potential sequence mislabelings in the database, each class C subtype has an inherently different level of discriminability according to the sequence characteristics.

Chapter 5

Grouping and visualization of mGlu subtype sequences

In this Chapter, we specifically focus on one of the most interesting and richly structured of class C subtypes, namely the mGlu receptors.

As described in section 2.1 from Chapter 2, the mGlu receptors, widely distributed throughout the CNS, play a relevant role in the regulation of cell excitability and synaptic transmission. They are divided into three groups (I, II, III) including eight subtypes distributed as follows: Group-I: mGlu1, mGlu5; Group-II: mGlu2, mGlu3; Group-III: mGlu4, mGlu6, mGlu7 and mGlu8.

5.1. Results and discussion

5.1.1. Subtyping mGlu receptors

In this section, we report the visualizations of data sets 3 and 4, previously described in section 2.1. Data set 3 is visualized using the MSA transformation (see section 2.2.2), while data set 4 is visualized using AAC and Digram transformations (see section 2.2.1). Firstly, data set 4 is visualized using the posterior mode projection-based KGTM. Then, the several subtypes of the corresponding mGlu receptors are displayed in figure 5.1. It reveals the distribution of the eight different subtypes extracted from the GPCRdb database.

It must be noted that in the primary data set, the mGlu7 subtype was absent. Instead, a new subtype denoted as *mGluLike* was present. It may be assumed that the *mGluLike* subtype includes receptors that are classified as mGlu receptors by GPCRdb, but without a fully true genetic adscription.

Strikingly, KGTM separates quite well each of the eight subtypes of mGlu receptors. Further detail of the mapped location of each subtype can be appreciated in the display of Fig. 5.1.

It is worth mentioning that the plot of mGlu subtypes displayed in Fig. 5.1 has been accomplished by the KGTM model previously obtained for class C. In other words, the KGTM model was not trained again on the mGlu subset; instead, the other types were made "silent" and sequences were labelled accordingly with their mGlu receptor subtype identity.

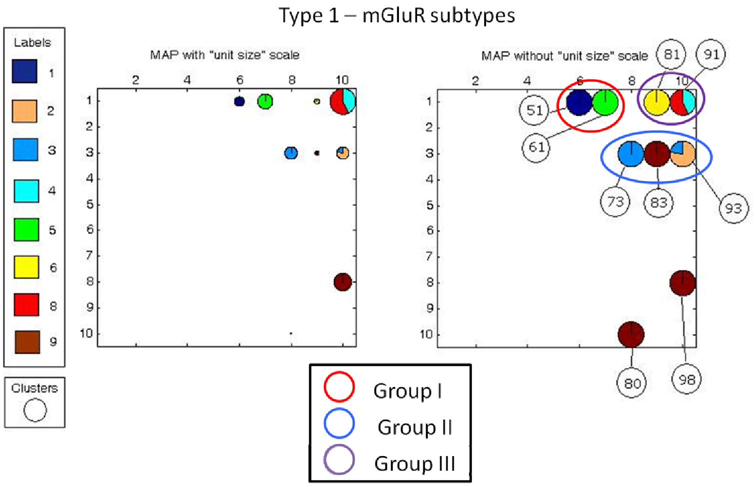


Figure 5.1: Mode projection of the mGlu receptor subtypes. Labels: 1: mGlu1, 2: mGlu2, 3: mGlu3, 4: mGlu4, 5: mGlu5, 6: mGlu6, 8: mGlu8, 9: *mGluLike*. The analysed data set has no mGlu7 subtype cases. There is a visible separation of the subtypes in three main groups, according to the amino acid sequence similarity, agonist pharmacology and the signal transduction pathways to which they couple: group I (mGlu1, mGlu5), group II (mGlu2, mGlu3, *mGluLike*) and group III (mGlu4, mGlu6, mGlu8)

According to this visualization, the mGlu receptor sequences of subtype 9 corresponding to *mGluLike*, assigned to cluster 83 are very homogeneous. They include the subtypes *mGluLike2* and *mGluLike3* and are well-located between *mGlu2* and *mGlu3*. On the other hand, the *mGluLike* groups assigned to clusters 80 and 98 are quite far from subtype 1- *mGlu receptors*, but very close to the subtypes 5 *(Vomeronasal)*, 6 *(Pheromone)* and 7 *(Odorant)* (See Fig. 5.2 for complete detail), taking into account their neighbourhood. This suggests that some GPCRdb assignments of *mGluLike* receptors to the mGlu group might be incorrect, and that they might in fact be smell sense receptors. This is only a hypothesis and would require further testing.

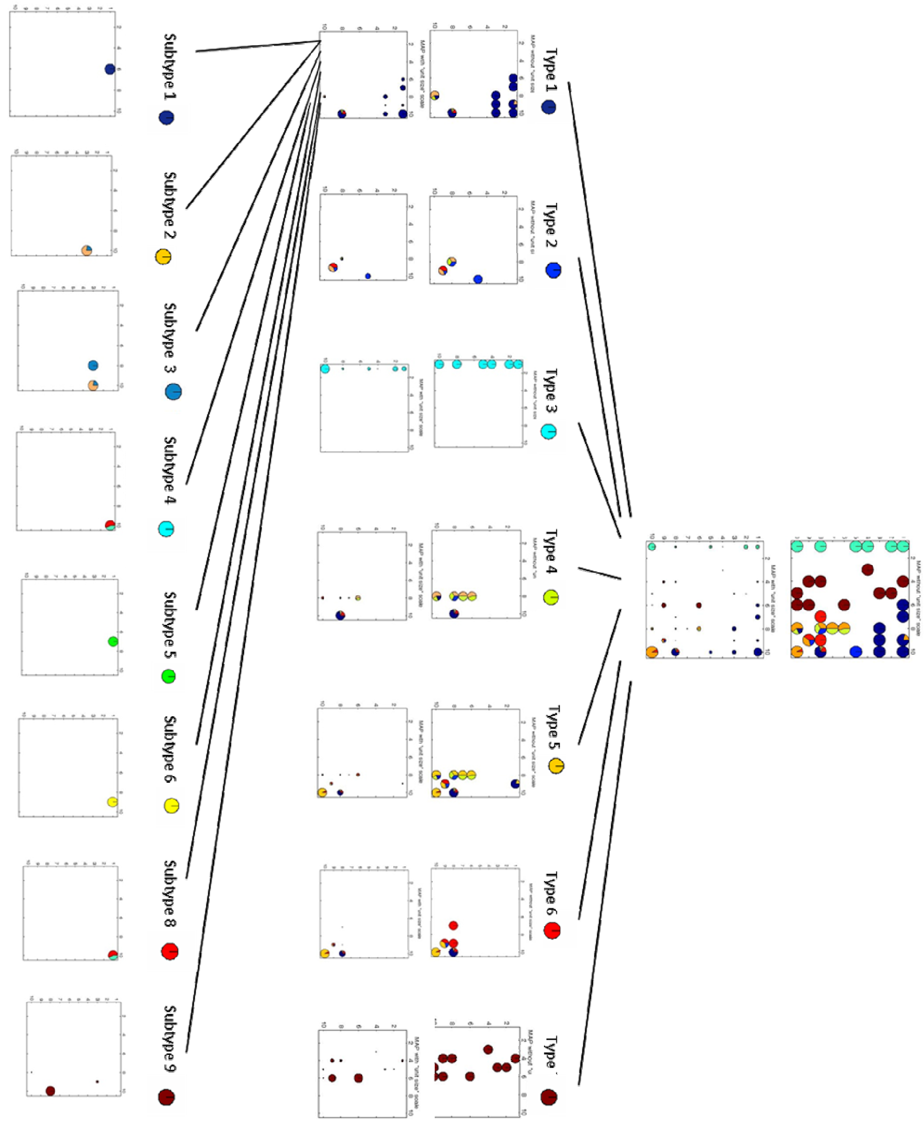


Figure 5.2: General hierarchical visualization of GPCR Family C types, including detailed subtyping of mGlu receptors.

Furthermore, data set 4 is visualized using the standard GTM-based posterior mean projection, the KGTM mode projection and the Radial PT. All GTM maps in the latent visualization space were created using a regular square grid of 15 × 15 points. The standard GTM visualization of the AAC-transformed mGluR sequences according to their *posterior mean projection* is shown in Fig. 5.3, whereas the similar visualization of the digram-transformed sequences is shown in Fig. 5.4.

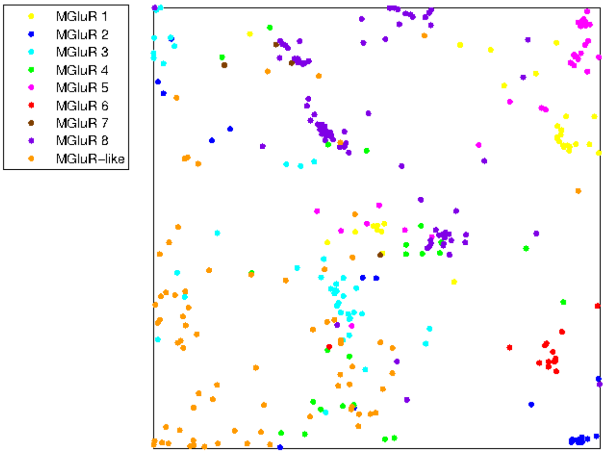


Figure 5.3: Visualization map of the standard GTM-based *posterior mean projection* of the mGluR AAC-transformed sequences. Different mGluR subtypes are identified by colour.

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 mGlu receptors is better and more intuitively represented by their *posterior mode projections* as shown in Fig. 5.5.

All GTM visualizations provide interesting insights about the inner grouping structure of mGlu receptors. The first overall finding is that most subtypes show a reasonable level of separation, but none of them avoids some level of subtype overlapping. Interestingly, most subtypes show clear inner structure themselves,

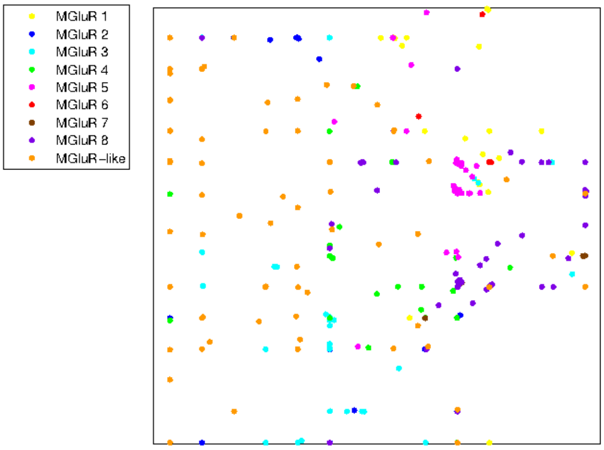


Figure 5.4: Visualization map of the standard GTM-based *posterior mean projection* of the mGluR digram-transformed sequences, as in previous figure.

which indicates that lower levels of sub-grouping might be worth investigating (for instance, through hierarchical clustering strategies). As an example, in Fig. 5.3, which corresponds to the AAC transformation, mGlu8 sequences are separated in at least four clearly delimited sub-groups. mGlu1, in turn, show at least one group around the center of the map and a second in its top-right corner. mGlu2, instead, seems to be mostly concentrated in the bottom-right corner.

The differences between the AAC sequence mapping and its *digram* counterpart in Fig. 5.4 are noticeable, although there are also clear coincidences, such as the neat separation of the rather heterogeneous mGluR-like sequences in the bottom-left quadrants of both maps, with the mGlu3 subtype located nearby. Overall, the differences indicate that the visual representation of these data and, consequently, the type of knowledge that can be inferred from it is at least partially dependent on the type of sequence transformation.

This is further corroborated by the KGTM visualization in Figure 5.5, using the *posterior mode projection*. The mapping differs in many ways from the

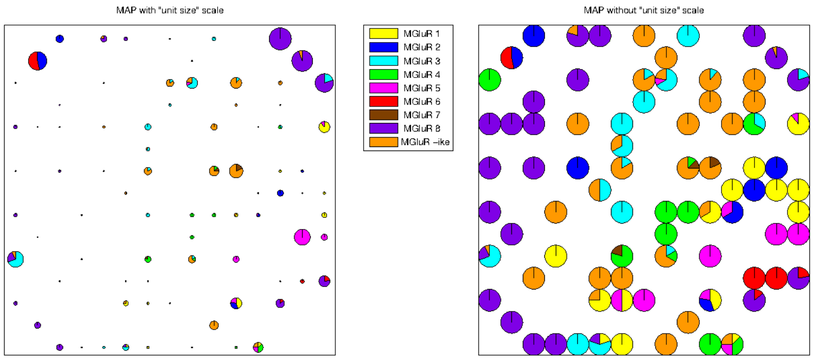


Figure 5.5: KGTM-based visualization of the mGlu receptor subtypes through their *posterior mode projection*. Left) Individual pie charts represent sequences assigned to a given latent point and their size is proportional to the ratio of sequences assigned to them by the model. Each portion of a chart corresponds to the percentage of sequences belonging to each mGlu. Right) The same map without sequence ratio size scaling, for better visualization. Subtype colouring as in Fig. 5.3

.

previous ones, although many characteristics remain consistent: the mGluR-like and mGlu3 are again located nearby, whereas the mGlu1, mGlu8 and others show evidence of inner sub-structure.

As stated in section 2.1, the eight main mGluR subtypes are commonly grouped into three categories: type I, including mGlu1 and 5; type II, including mGlu2 and 3; and type 3, including mGlu4, 6, 7 and 8. The visualizations in Figures 5.3, 5.4 and 5.5 provide only partial support to these categories. Type I seems quite coherent in all representations, regardless data transformation. Type II, instead, is not clearly homogeneous according to any of them. Similarly, limited homogeneity is observed in the four subtypes of Type III for all transformations.

In order to quantify the level of subtype overlapping to support the preliminary visual impressions, we again used the previously defined entropy-based measures. The results for the standard GTM representation of the AAC- and *Digram*-transformed sequences, as well as for the KGTM MSA-transformed sequences are summarized in Table 5.1.

Table 5.1: Entropies per mGluR and mGluR-like subtype, together with total entropy for each of the data transformations and GTM variants.

mGluR subtype	GTM map entropy		
	GTM-AAC	GTM-digram	KGTM
mGluR1	0.35	0.77	0.50
mGluR2	0.41	0.55	0.65
mGluR3	0.69	0.53	0.51
mGluR4	0.65	0.59	0.47
mGluR5	0.55	0.50	0.53
mGluR6	0.19	0.80	0.57
mGluR7	0.41	0.69	0.64
mGluR8	0.33	0.48	0.27
Like	0.48	0.35	0.50
Total Entropy	0.31	0.37	0.33

Beyond the qualitative appreciation of similarities and dissimilarities between sequence sub-groups, the entropy measure can provide the analyst with an at least overall measure of subtype location specificity, which should be a clear clue about potential subtype discriminability in a classification setting. The results summarized in Table 5.1 are quite telling. First, because the overall entropy is not too dissimilar between sequence transformations; despite this, the transformation yielding lowest entropy (and, therefore, highest level of subtype discrimination) is, unexpectedly, the simplest one: AAC, which does not even consider ordering in the receptor 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: the entropy levels of several subtypes differ quite widely between transformations (although in subtypes with a low number of sequences, such as mGluR6, these differences should be considered with caution), while the entropy level ranking differs completely.

Complementary grouping-based visualization is provided by Treevolution PTs. Figures 5.6, 5.7 and 5.8 illustrate their use. In Fig. 5.6, we see mGluRs in the context of the complete class C GPCR set (1,510 sequences). In Figures 5.7 and 5.8, we show the Type III subtypes (mGluR4, mGluR6, mGluR7 and mGluR8) again in the context of the complete class C GPCR set. For the sake of brevit

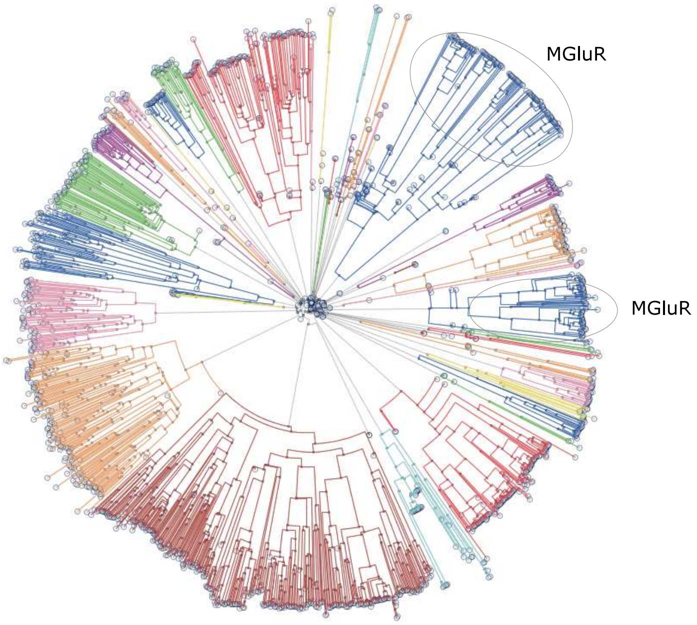


Figure 5.6: PT of the mGluR sequences in the context of class C GPCRs. Each terminal node in the exterior of the hierarchical radial display corresponds to an individual sequence.

The visualization of the mGluR using a radial PT, in Fig. 5.6, is a good complement for the GTM-based visualization. It shows how the mGluR subtypes (in blue, top right) are neatly distributed in two differentiated sectors of the tree, which is coloured according to its common ancestors and the depth level. This implies that, in some cases, mGluR inter-subtype similarity is lower than the similarity between mGluR and other class C GPCR subtypes.

Figures 5.7 and 5.8, which represent through PTs the mGluR subtypes corresponding to *type III*, provide some explanation for this: All type III sequences reside in only one of this sectors. That is, the PT models them as having

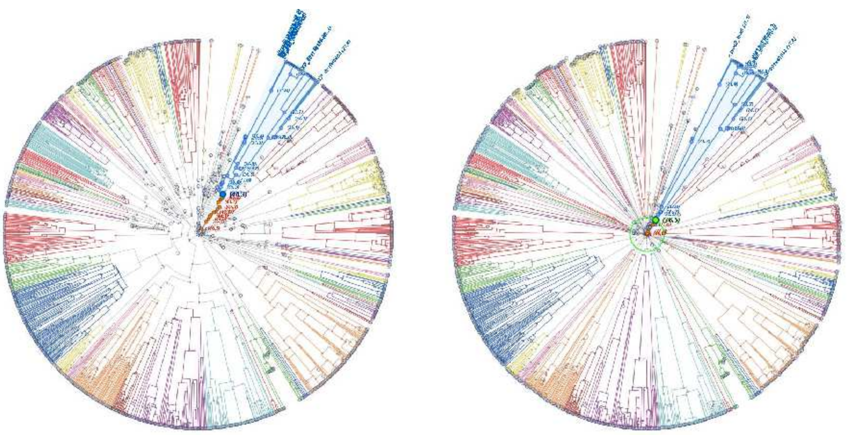


Figure 5.7: PT highlighting mGlu4 and mGlu6, part of Type III, in the context of class C GPCRs.

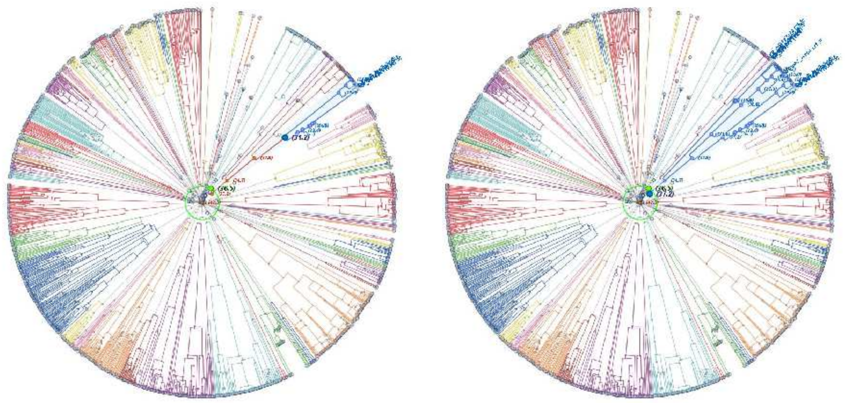


Figure 5.8: PT highlighting mGlu7 and mGlu8, part of Type III, in the context of class C GPCRs.

common ancestors with a high degree of dependency. Curiously, mGlu4 and mGlu6, on one side, and mGlu7 and mGlu8, on the other, overlap extensively. Note that this is partially in contradiction with the visualizations shown in previous figures, in which overlapping between mGlu4 and mGlu6 is hardly observed, whereas overlapping between mGlu7 and mGlu8 can be observed in the standard GTM visualizations based on the AAC and *Digram* transformations.

Chapter 6

Analysis and Visualization of Classification Errors in Class C GPCRs

Previous research employing supervised and semi-supervised learning methods for the classification of the different subtypes of class C GPCRs has revealed the existence of a *soft* upper boundary on the accuracy that can be achieved in their discrimination from the unaligned transformation of their sequences [33, 106].

Given that the target of this thesis is the exploration of data sequences using unsupervised learning methods oriented towards sequence visualization, we investigate the characteristics of such boundary by focusing on those sequences that were consistently misclassified using supervised methods.

These sequences are visualized, again, using nonlinear dimensionality reduction and PTs, 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 the initial proof of concept for the proposed method.

Given the exploratory goal of this experiments, we perform them using the 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 [33] 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 [106] yielded best results in the area of 88%. A detailed review about this type of classification can be found in [166].

To investigate this classification bound, we present in this Chapter a method that combines GPCR classification with MVD visualization using the unaligned transformed sequences as a starting point.

Firstly, we 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 GTM which is described in section 3.1.1. 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

Table 6.1: Number of class C misclassified sequences, listed by subtype.

Subtype	Number of misclassified sequences
mGlu	16
CaS	5
GABA-B	8
VN	46
Ph	48
Od	35
Ta	5
Total	163

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 the PT technique detailed in section 3.3.

6.1. Results and discussion

6.1.1. List of class C misclassified cases

Experiments were performed for data set 2, which corresponds to the class C family subtypes listed in section 2.1. A batch of previous supervised classification experiments using SVMs were used as the starting point [106]. Such experiments involved an iterative 5-fold cross-validation (CV) process, splitting the data set 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 6.1) in the sense that these sequences were most often classified to the same wrong class.

Table 6.2: Misclassified mGlu sequences. List of the 16 misclassified mGlu, including their GPCRdb identifier (ID), their class as predicted by SVM and their sequence name.

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

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 6.4: 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	Sequence 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 6.5: Misclassified Ta sequences. List of the 5 misclassified Ta, including their GPCRdb identifier (ID), their class as predicted by SVM and their sequence name.

Sequence ID	Predicted subtype	Sequence 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 6.6: Misclassified VN, Ph and Od sequences. Summary list of the largest groups of misclassifications.

Sequence ID	Predicted subtype	Sequence name
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

Tables 6.2 to 6.5 list all the misclassified sequences from mGlu, CaS, GABA-B and Ta subtypes in detail. The characteristics of the far more abundant misclassifications of VN, Ph and Od subtypes are summarily reported in Table 6.6 and reported in detail in tables 6.7, 6.8 and 6.9.

N^o	Sequence ID	Predicted subtype	Sequence name
1	683	mGlu	XP_002937102
2	749	mGlu	XP_002941318
3	753	mGlu	XP_002941322
4	756	mGlu	XP_002941777
5	764	mGlu	XP_002942628
6	784	mGlu	XP_002943694
7	851	mGlu	NP_001093066
8	691	CaS	XP_002937455
9	738	CaS	XP_002941226
10	676	Ph	XP_002936197
11	677	Ph	XP_002936218
12	681	Ph	XP_002936334
13	686	Ph	XP_002937448
14	699	Ph	XP_002938198
15	723	Ph	XP_002940457
16	724	Ph	XP_002940458
17	725	Ph	XP_002940476
18	729	Ph	XP_002940322
19	736	Ph	XP_002941221
20	740	Ph	XP_002941228
21	769	Ph	XP_002943507
22	772	Ph	XP_002942710
23	778	Ph	XP_002943137

24	780	Ph	XP_002943139
25	834	Ph	XP_002721975
26	840	Ph	NP_001093048
27	852	Ph	NP_001093039
28	857	Ph	a0t300_danre
29	860	Ph	NP_001092974
30	867	Ph	q501x9_danre
31	868	Ph	a3kqh4_danre
32	869	Ph	NP_001098650
33	875	Ph	NP_001098007
34	876	Ph	NP_001092975
35	892	Ph	NP_001093005
36	909	Ph	XP_001516454
37	927	Ph	NP_001098528
38	944	Ph	NP_001093129
39	951	Ph	NP_001098526
40	621	Od	q8bid7_mouse
41	747	Od	XP_002941317
42	773	Od	XP_002942711
43	884	Od	o70411_rat
44	885	Od	q8tdu1_human
45	886	Od	o70413_rat
46	900	Od	XP_917917
	Total VN		46 sequences

Table 6.7: SVM misclassified sequences labeled as Vomeronasal (VN) in GPCRdb. The second column provides the sequence ID, the third column is the subtype predicted by SVM, while the fourth column is the sequence name.

N^{o}	Sequence ID	Predicted subtype	Sequence name

1	952	mGlu	*a7sdg9_nemve*
2	953	mGlu	*a7s1x6_nemve*
3	954	mGlu	*a7s0d2_nemve*
4	956	mGlu	*b3s609_triad*
5	957	mGlu	*XP*_001494824
6	958	mGlu	*XP*_002731604
7	959	mGlu	*XP*_002732067
8	961	mGlu	*XP*_002935674
9	1090	mGlu	*XP*_002937659
10	1097	mGlu	*XP*_002940462
11	1117	mGlu	*XP*_002940343
12	1164	mGlu	*XP*_002943384
13	1196	mGlu	*XP*_001505324
14	1282	mGlu	*a8e7k1_danre*
15	1310	mGlu	*XP*_684341
16	1321	mGlu	*XP*_001509767
17	1324	mGlu	*q9pwe1_ictpu*
18	1325	mGlu	*b0uyj3_danre*
19	1332	mGlu	*XP*_001521075
20	955	GABA-B	*b3s157_triad*
21	1055	GABA-B	*q4spr3_tetng*
22	1104	GABA-B	*XP*_002939765
23	1176	GABA-B	*XP*_002942720
24	960	VN	*XP*_002933303
25	1073	VN	*XP*_002935803
26	1078	VN	*XP*_002936336
27	1128	VN	*XP*_002941492
28	1149	VN	*XP*_002941355

29	1151	VN	XP_002941357
30	1155	VN	XP_002942464
31	1158	VN	XP_002941770
32	1171	VN	XP_002942717
33	1182	VN	XP_002943278
34	1191	VN	XP_001368172
35	1266	VN	XP_002723672
36	1270	VN	NP_001093018
37	1272	VN	NP_001093020
38	1273	VN	NP_001093022
39	1274	VN	NP_001093016
40	1275	VN	NP_001093017
41	1291	VN	o35272_rat
42	1299	VN	XP_002723938
43	1302	VN	XP_002936172
44	1328	VN	NP_001093040
45	1334	VN	XP_001516991
46	1329	Od	XP_002944635
47	1330	Od	XP_696754
48	1331	Od	XP_001075542
Total Ph			48 sequences

Table 6.8: SVM misclassified sequences labeled as Pheromone (Ph) in GPCRdb. The second column provides the sequence ID, the third column is the subtype predicted by SVM, while the fourth column is the sequence name.

Table 6.9: SVM misclassified sequences labeled as Odorant (Od) in GPCRdb. The second column provides the sequence ID, the third column is the subtype predicted by SVM, while the fourth column is the sequence name.

N°	Sequence ID	Predicted subtype	Sequence name
1	1405	mGlu	*XP_001520670*
2	1409	mGlu	*b3rud8_triad*
3	1414	mGlu	*XP_002936183*
4	1427	mGlu	*XP_002941773*
5	1399	VN	*gpc6a_human*
6	1410	VN	*d1lwx7_sacko*
7	1411	VN	*XP_002727501*
8	1412	VN	*XP_002933716*
9	1413	VN	*XP_002936177*
10	1419	VN	*XP_002940566*
11	1421	VN	*XP_002940324*
12	1422	VN	*XP_002940329*
13	1423	VN	*XP_002941570*
14	1424	VN	*XP_002941571*
15	1425	VN	*XP_002941572*
16	1426	VN	*XP_002942058*
17	1428	VN	*XP_002941794*
18	1431	VN	*XP_002943912*
19	1344	Ph	*b0s550_danre*
20	1345	Ph	*q5i5c7_9tele*
21	1346	Ph	*a3qjy1_danre*
22	1347	Ph	*a0t301_danre*
23	1348	Ph	*a3qjy3_danre*
24	1350	Ph	*a8e7t9_danre*
25	1351	Ph	*XP_001332644*
26	1353	Ph	*XP_001332817*
27	1356	Ph	*a3kql8_danre*
28	1380	Ph	*XP_001332729*
29	1386	Ph	*q6unx3_ictpu*
30	1404	Ph	*XP_001518611*
31	1416	Ph	*XP_002937663*
32	1417	Ph	*XP_002939763*
33	1418	Ph	*XP_002940477*
34	1435	Ph	*XP_002944421*
35	1445	Ph	*a3qjy2_danre*
Total Od			35 sequences

6.1.2. Visualization of misclassified sequences using GTM

The data set 2 from section 2.1 was preprocessed with the AAC transformation and then visualized using the posterior mean projection of GTM, as described in previous Chapters. This global GTM visualization map is displayed in Fig. 6.1. 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 6.2 to 6.8. In each of these maps, the misclassified sequences are individually identified using the sequence ID.

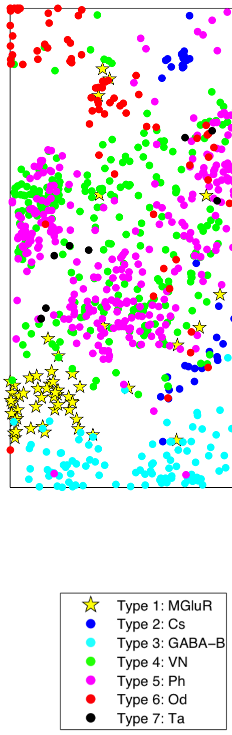


Figure 6.1: GTM posterior mean projection of the data set and list of corresponding labels. Visualization of all 1,510 sequences. Each color corresponds to a GPCR class C subtype.

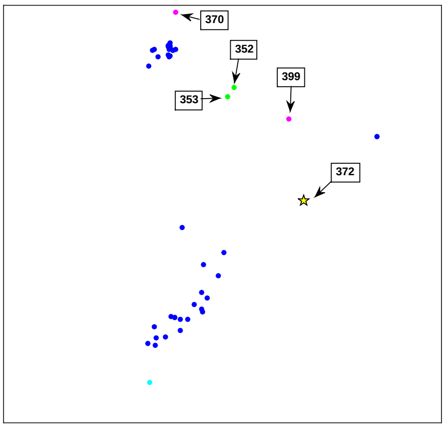


Figure 6.3: CaS GTM posterior mean projection. Visualization of CaS sequences. Representation as in Figure 6.2. Cases labeled with their ID from Table 6.3.

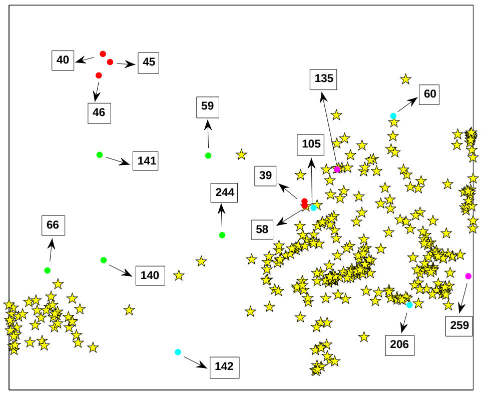


Figure 6.2: mGlu GTM posterior mean projection. Visualization of mGlu sequences. Cases incorrectly classified by SVM are represented with the colors of their predicted subtypes. Cases labeled with their ID from Table 6.2.

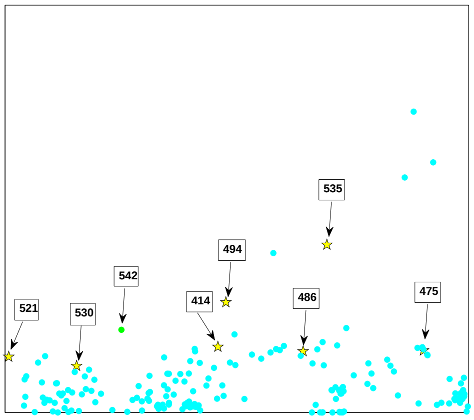


Figure 6.4: GABA-B GTM posterior mean projection. Visualization of GABA-B sequences. Representation as in Figure 6.2. Cases labeled with their ID from Table 6.4.

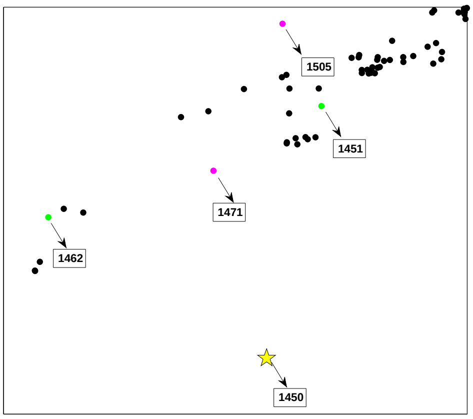


Figure 6.5: Ta GTM posterior mean projection. Visualization of Ta sequences. Representation as in Figure 6.2. Cases labeled with their ID from Table 6.5.

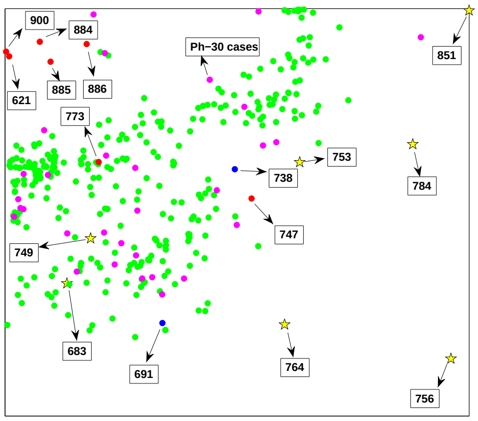


Figure 6.6: VN GTM posterior mean projection. Visualization of VN sequences. Representation as in Figure 6.2. Cases labeled with their ID from Table 6.7. Note that the 30 Ph misclassified cases are not individually labeled.

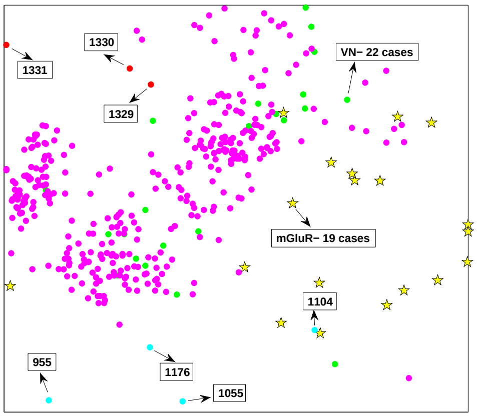


Figure 6.7: Ph GTM posterior mean projection. Visualization of Ph sequences. Representation as in Figure 6.2. Note that the 22 VN and 19 mGlu misclassified cases are not individually labeled.

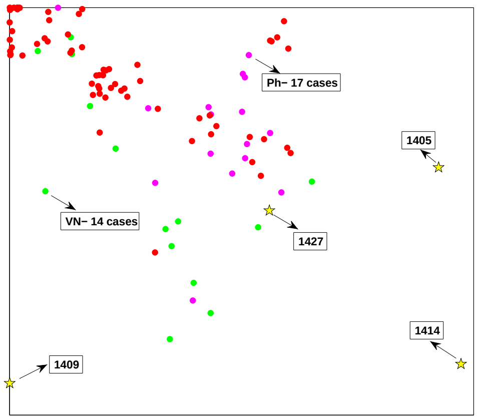


Figure 6.8: Od GTM posterior mean projection. Visualization of Od sequences. Representation as in Figure 6.2. Note that the 14 VN and 17 Ph misclassified cases are not individually labeled.

It is clear from the GTM visualization of the complete set of transformed class C GPCR sequences (Fig. 6.1) 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 6.2 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 6.8). 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 [111]. 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 (a database curator, for instance) 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 6.2), for instance, the pair $a8dz71\_danre$ and $a8dz72\_danre$, according to the UniProt [I] 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 [II], $q5i5d4\_9tele$ 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 [III], whereas the latter was "removed as a result of standard genome annotation processing" from NCBI [IV]. Finally, $XP\_002123664$, misclassified as an mGlu, was also "removed as a

[I] http://www.uniprot.org/uniprot/A8DZ72

[II] http://www.ebi.ac.uk/ena

[III] http://www.ncbi.nlm.nih.gov/protein/XP_001515899.2

[IV] http://www.ncbi.nlm.nih.gov/protein/XP_002740613 result of standard genome annotation processing" from NCBI [V], 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 [VI] 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.

6.1.3. Visualization of misclassified cases with Radial PT

A PT of the complete set of 1,510 sequences was created using Treevolution software. It is shown in Fig. 6.9 and it is used here 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.

[V] `http://www.ncbi.nlm.nih.gov/protein/XP_002123664`

[VI] `http://www.ncbi.nlm.nih.gov/protein/XP_002738008`

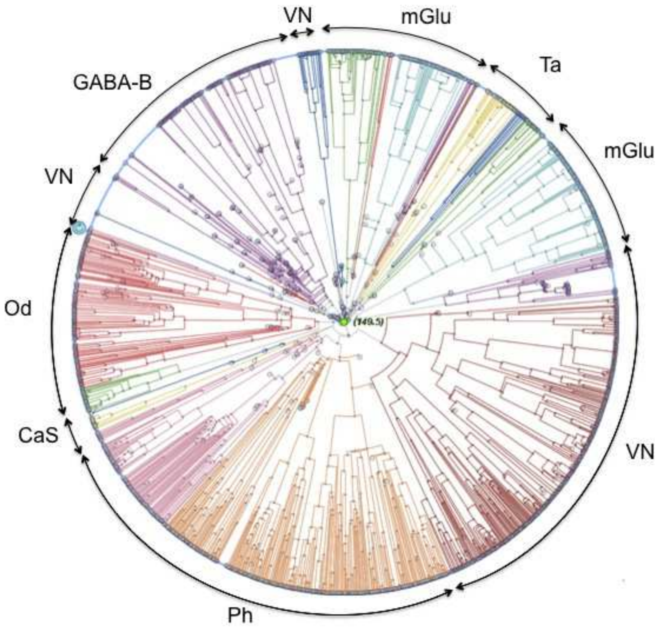


Figure 6.9: Treevolution radial PT plot of the 1,510 GPCRs. Each branch corresponds to one GPCR sequence. Two separated mGlu sections can be identified, as well as three consecutive CaS sections; a single GABA-B section; three separate VN ones; two consecutive groups of Ph; two of Od and three consecutive 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.

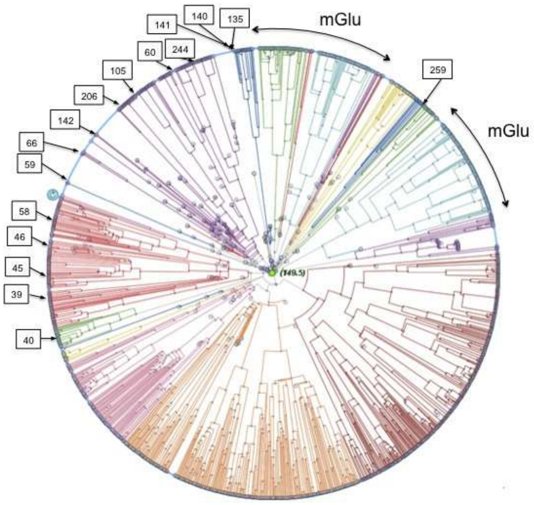


Figure 6.10: Radial PT plot for mGlu misclassified cases.

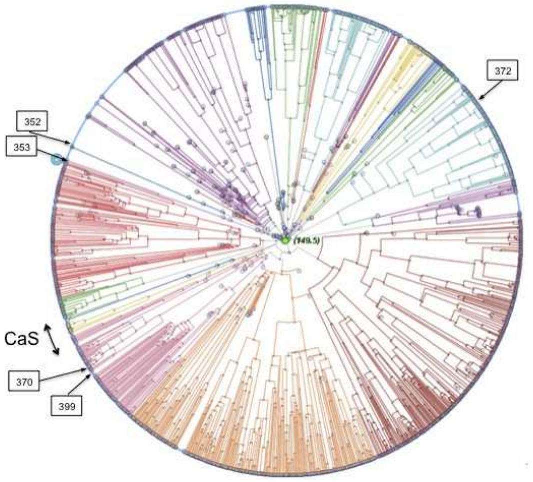


Figure 6.11: Radial PT plot for CaS misclassified cases.

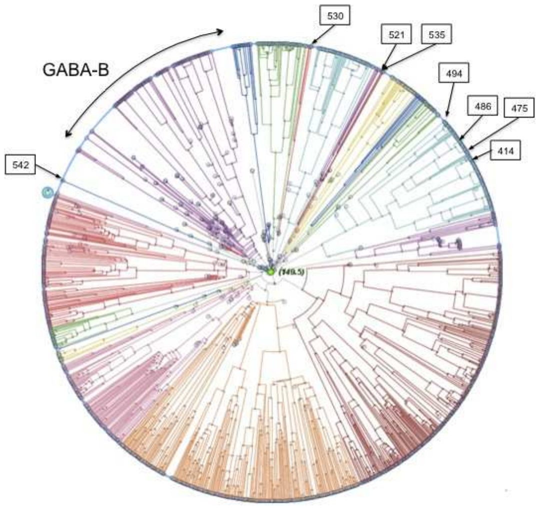


Figure 6.12: Radial PT plot for GABA-B misclassified cases.

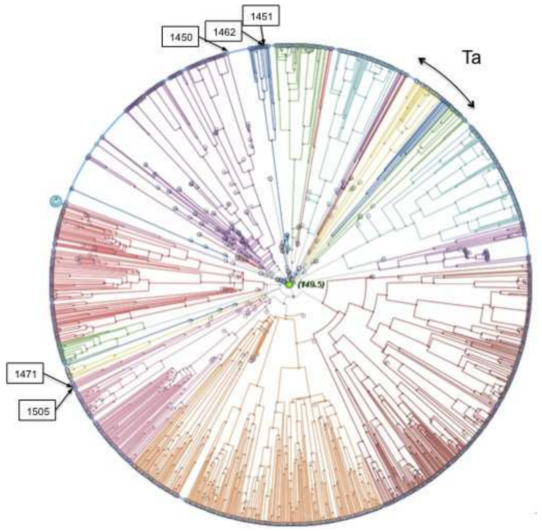


Figure 6.13: Radial PT plot for Ta misclassified cases.

Fig. 6.9 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 6.10 to 6.13 for, in turn, subtypes mGlu, CaS, GABA-B and Ta are quite consistent with those shown in GTM Figures 6.2 to 6.5. 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 Fig. 6.10, 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 Fig. 6.10 are located in the corresponding areas of their predicted subtypes.

The results visualized in Figures 6.11, 6.12 and 6.13 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.

6.1.4. The effect of sequence size on class C GPCR subtype classification

The experiments applied to data set 2 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 section provides some evidence to support this expectation. For that, we show in Fig. 6.14, 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. The vertical axis does not reflect absolute numbers but relative frequencies, so that both can straightforwardly be compared.

Note that both distributions of lengths have very similar shapes, with the highest frequencies for misclassified sequences located at the same range of lengths as the highest frequencies for the complete sample. There are differences between the frequencies of misclassification at certain lengths and the corresponding frequencies of those lengths in the overall sequence population, but never too substantial. Misclassification is somehow higher in sequences of length lower than 700 AAs, but, interestingly, the frequency of misclassification at the lowest lengths (under 300 AAs) is far lower than the frequency of those lengths in the complete data set. Similarly, misclassification is somehow higher in sequences of lengths over 1,700 AAs, but the frequency of misclassification at lengths between 1,100 and 1,300 AAs is lower than the frequency of those lengths in the complete data set. It could thus be safely stated that the sequence length has, at most, a moderate overall effect on the misclassification process.

A quantitative measure of this effect can be calculated through an approximation of the conditional probability of misclassification, given the length of the sequences. This conditional probability, following Bayes theorem, would take the form:

$$P(mc|l) = P(l|mc)P(mc)/P(l) \tag{6.1}$$

Table 6.10: A group of 7 sequences of length over 1,500 AAs.

90	$XP\_002157920$	mGlu	1712
234	$q8nha9\_human$	mGlu	1520
535	$XP\_002738008$	GABA-B	1995
714	$XP\_002937835$	VN	1656
756	$XP\_002941777$	VN	1738
1108	$XP\_002940475$	Ph	1768
1176	$XP\_002942720$	Ph	1869

where $P(l|mc)$ and $P(l)$ could be approximated by the previously shown bar charts and $P(mc) = 163/1510 = 0.108$.

A graphical representation of this measure, with as a horizontal line indicating no divergences between the distributions of misclassified sequences and the total sample at a given interval of lengths, can be found in Figure 6.15.

Note that the values for sequences of lengths beyond 1,400 AAs have little statistical significance. To understand this, you have to bear in mind that the vector of absolute numbers of sequences from which the histogram of the complete sample was built is: (16, 30, 20, 41, 86, 236, 708, 183, 53, 49, 45, 28, 8, 1, 1, 3, 1, 1), while the corresponding vector for the misclassifications is: (1, 8, 7, 8, 15, 25, 60, 16, 7, 4, 2, 4, 3, 0, 0, 1, 1, 1). That is, only 15 sequences (less than 1% of the total) have lengths of more than 1,400 AAs. In particular, the 7 sequences of length over 1,500 AAs are detailed in Table 6.10.

Out of these, only the GABA-B and the two pheromones were misclassified. This graphical representation corroborates the conclusions that were drawn from the histograms.

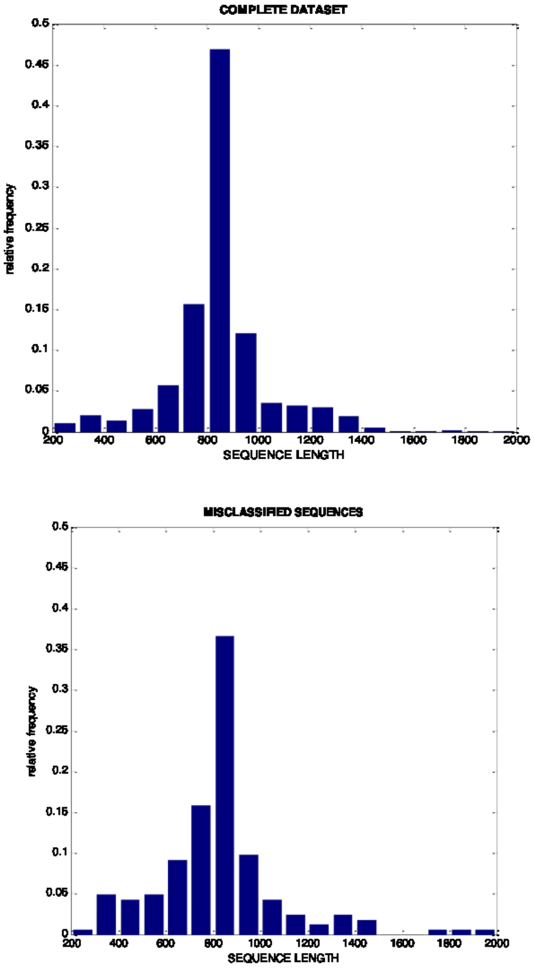


Figure 6.14: Comparison of sequence length for complete data set and for misclassified sequences.

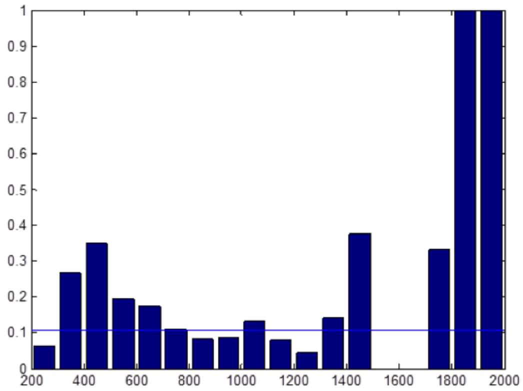


Figure 6.15: Comparison of sequence length for complete data set and for misclassified sequences.

Chapter 7

Conclusions and Future Work

The use of Machine Learning (ML) methods for the analysis of multivariate data (MVD) is a process of knowledge extraction. Acquiring knowledge is not the same as obtaining results, though, regardless their quality. Knowledge extraction from MVD requires results to be interpretable by the analyst [198]. Interpretability is a bottleneck particularly for nonlinear ML techniques, which means that a trade-off between flexibility and performance, on one side, and interpretability, on the other, must be achieved.

This bottleneck is important in computational biology [21], biomedicine [146] and bioinformatics. The study carried out in this Thesis follows an exploratory approach to knowledge extraction in which MVD visualization is the key component. Visualization becomes relevant for the analysis of high-dimensional data, as it opens a door to inductive reasoning [196] and, thus, interpretability.

We have addressed a bioinformatics problem: the analysis of a database of G-protein-coupled receptors (GPCR) from their amino acid (AA) sequences. This is part of an effort to investigate the extent to which their subtypes can be discriminated. Such discrimination has been attempted at great detail [63]. We focused on the class C family of GPCRs, which is characterized, first, by the fact that the full 3-D structure has not yet been solved completely for any of its members (only partial domains 7TM, ECD and VFT were crystallized [112], [193], [139], [42], [67], [213], [205]), restricting the investigation of their functionality on their primary structure, that is, their AA sequence, and, second, by its great interest in pharmacology due to its increasing therapeutic prospects. The visualization of the high-dimensional class C GPCR sequences was carried out here using different versions of Generative Topographic Mapping (GTM [19]), a model that has successfully been applied in biomedicine and bioinformatics (Chapters 4, 5 and 6). These GTM versions have been applied, in turn, to different transformations of the sequences, with and without sequence alignment (Chapter 3). The analysis focused not on subtype discriminability but, instead, on subtype overlapping and its hypothetical consistency over different sequence transformations. This could be a preliminary step for the future investigation of heterodimerization in class C GPCR subtypes.

7.1. Overview of the main contributions and conclusions of the research

My research has resulted in the following general contributions to the computational intelligence analysis of GPCRs:

- I have described novel and improved existing CI-based methods, including the definition of a adapted kernel method for proteomics. These methods have been applied to solve a number of problems and the conclusions of this research are summarized in the following subsections.

- I have published this applied and theoretical research in international journal and conferences, mainly targeting those at the interface between data science and bioinformatics.

7.1.1. Visualization of Class C GPCR types

In Chapter 3, we have described a kernel method of the manifold learning family that is capable of simultaneously revealing the grouping structure of GPCRs while making the intuitive visualization of such structure possible.

Our results are consistent with early classification studies using other techniques such as Hidden Markov Models, thereby validating the present methodology. Importantly, the method herein presented reveals mixing between some receptor subclasses, suggesting its possible applicability to the study of heterodimerization between receptors. Receptor heterodimerization has been confirmed experimentally for a number of receptors. This finding paves the way for new strategies in drug discovery research providing a conceptual framework for the rational combination of drugs. KGTM may help in the exploration of receptors susceptible of heterodimerization and thus be useful in the quest of more potent and safer drugs.

Phylogenetic trees, detailed in Chapter 3, and applied to our data to complement the visualization (Chapters 4, 5 and 6), are a widely used graphical tool in the field of proteomics. They illustrate the probability that two sequences are more closely evolutionary related to each other than to a third one. The reported results show that the unsupervised mapping of GPCR sequences yielded by KGTM closely resembles the corresponding phylogenetic trees to a great deal of detail. This corroborates that KTGM could be used as a complement to the phylogenetic tools, as it provides users with a very detailed while easy to interpret visual grouping of the sequences that is fully consistent with the more complex PT representation.

In the absence of knowledge about their tertiary structure, class C GPCRs do often have to be investigated according to the primary structure of their AA sequence. Different techniques have been developed that either deal with the complete unaligned sequence or its aligned transformation. We have analyzed, helped by NLDR visualization techniques that handle the GPCR sequences as MVD, the heterogeneous way in which the seven known subtypes of class C overlap.

Overlapping indicates (partial) similarities between subtypes, whose relation with processes of GPCR heteromerization could be the matter of future research. Considering that heteromerization involves physical interaction between some but not all of the 7 helices of the transmembrane domain of GPCRs, a more detailed subtype overlapping analysis might be required. Results add interpretability to an otherwise complex problem and show that the level and quality of this overlapping depends on the data transformation and NLDR modelling technique, although consistent enough to reach conclusions of interest to the analyst.

7.1.2. Visualization of Class C GCPR mGluR subtypes

Metabotropic glutamate receptors are the target of intensive research due to their impact on the design of drugs for a wide array of pathologies. They also have been implicated in long-term potentiation and in learning and memory formation. The discovery of metabotropic glutamate receptors (mGluRs) 10 years ago, allowed more thorough insights in their functions. Since then, much effort has been centered on the cloning and characterization of the different mGluRs subtypes and the elucidation of their physiological function in multiple regions of the brain. As detailed in section 2.1 from Chapter 2, eight subtypes of mGluRs have been divided into three groups according to their sequence homology, pharmacological characterization, and coupling to second messenger pathways.

Group I mGluRs (mGluR1 and mGluR5), the only group which have partially solved crystal structures ([205], [42]), are predominantly localized at the membrane of the post-synaptic cells and regulate variety of physiological functions. These receptors have been implicated in various forms of synaptic plasticity including learning and memory as well as in various neuropsychiatric disorders [156], [98]. The involvement of different mGluR groups and subtypes in particular physiological circuits and functions such as hippocampal synaptic plasticity and learning is still a matter of controversial debate. Thus, given the lack of knowledge about their complete 3-D structure, their study focuses on primary sequential information. The data sets 4 and 5, consisting of mGluRs, were studied in Chapter 5 using a machine learning unsupervised approach based on NLDR and bioinformatics methods for the exploratory analysis of mGluRs through visualization. The experimental results provide evidence of the very rich inner substructure of these class C GPCRs, which not always conforms to their existing subtype labels. They also indicate that this substructure differs considerably depending on the data transformation method applied. In any case, such results caution that mGluR subtype classification is likely to be challenging and, importantly, that the own labelling of mGluR sequences, even if from curated databases, should be carefully investigated.

Thus, our proposed future research is expanding the proposed approach to the rest of class C GPCR subtypes to a deeper level of classification, as well as to more elaborate transformations and to selected part of the sequences, such as the extra-cellular domains.

7.1.3. Visualization of Misclassified Class C GPCRs

In Chapter 6 of this Thesis, we analyzed and visualized classification errors in Class C GPCR primary sequences. The study focuses on unsupervised data visualization using GTM, starting from previous supervised classification experiments reported in [106].

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 [57], [28].

Then, 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. The method adresses an exploratory visualization, using nonlinear dimensionality reduction techniques, of GPCRs previously shown to be misclassified using supervised (SVM) techniques. This is meant to be a proof of concept for a method that combines supervised classification and unsupervised grouping and visualization to assist in the task of GPCR database quality control. They were subsequently characterized against the rest of the data and, particularly, against the rest of cases of their own subtype. 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 [53], built using HMMs 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.

We argue that the exhaustive experimental setting used for the SVMs provides us not just with misclassifications, but also with solid evidence of misclassifications not just occurring due to specific quirks of the model, but due to the specificities of the data. A PT instead, will only provide us with a similarity-based grouping that does not necessarily reflect a functional relationship between the sequences and their labels. In other words, in the setting of our work we should only use PTs as a way to either confirm or challenge the misclassifications as inferred by SVMs and visualized by GTM.

Moreover, even if PTs, in this case, confirm the results that we obtain with the combination of SVMs and GTM, the latter model still provides us with information that PTs do not provide us with: The relative locations and interdistances between sequences in a latent/visualization metric space. In comparison, it also fair to say that GTMs lack the evolutionary hierarchical structure that PTs provide.

In short, PTs, on one side, and SVM+GTM, on the other, do not compete to give the same answer to the same question; they "collaborate" by confirming each other results from different perspectives and providing different elements of knowledge. 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.

7.1.4. Visualization of class C GPCRs parcial domains and Discrimination measures

Class C GPCR subtype separability was investigated in section 4.2 both through qualitative visual exploration using manifold learning methods and through several quantitative measures. We were specifically interested in the investigation of the different roles that might be played by the different domains of the Class C GPCR sequences on the subtype discrimination capabilities. In particular, the adequacy of the separate use of the extra-cellular N-Terminus and the transmembrane 7TM domain for subtype discrimination purposes was assessed. Results indicate that the use of KGTM with MSA provides the best discrimination, and that overall discriminability significantly decreases when only the N-Terminus or 7TM domains are used, but with partially mixed patterns for different Class C subtypes depending on the data transformation and manifold learning model used.

7.2. An outlook of future research

Much effort is made in the creation and maintenance of international, publicly-available curated databases for the bioinformatics domain. In protein databases, one of the tasks this effort entails is the labeling of proteins according to a not-fully standardized taxonomy of family and subtype assignments. This labeling process, which is often model-based, falls within the remit of classification problems.

This Thesis has focused on the analysis of GPCR cell membrane proteins of Class C, which have of late created great expectations in pharmacology as targets of drug design. From a discriminatory classification viewpoint, they have a heterogeneous subtype structure and, because of the absence of crystal structures including all the domains of any of these receptors, the investigation on their primary sequential structures can be of great help.

The classification of class C GPCRs from their transformed unaligned primary sequences seems to have a limiting classification threshold. Thus, in Chapter 6, 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. Therefore, 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 [63].

From the results provided in this study, future research could deal with the following question: to what extent inter-subtype sequence similarity is related to heterodimerization?

The concept of heterodimerization of two types of GPCRs was proposed in the early 1980's [1]. It represents an unexpected mechanism for GPCR regulation and function because the functioning of a heterodimer can be very different to that of each of its subunits, thereby providing a potentially novel target in pharmacology. Heterodimerization seems to be selective, so that GPCRs will interact with one type of receptors, but not with others [71, 133].

The *Calcium Sensing* receptors have mGluR1 and mGluR5 as heterodimer partners. This functional positive interaction is reported in [60]. Furthermore, class C Taste and GABA-B receptors are reported as canonical GPCR heterodimers whose selectivity lies in the Venus flytrap domain (VFD), where the active site for ligand binding is located [71]. It would be interesting to analyze the heterogeneous *mGlu* overlapping and the high *Calcium Sensing* entropy reported in the previous section in the light of these recently reported results. Since GPCRs are key drug targets in the treatment of different diseases, understanding the specificity and physiological significance of GPCR heteromerization may lead to insights that will impact the development of future therapeutics [161].

Bibliography

[1] Agnati L.F., Fuxe K., Zoli M., Rondanini C., and Ogren S.O. (1982) New vistas on synaptic plasticity: the receptor mosaic hypothesis of the engram. *Medical Biology*, 60: 183-190.

[2] Alexander, S.P., Davenport, A.P., Kelly, E., Marrion, N., Peters, J.A., Benson, H.E., Faccenda, E., Pawson, A.J., Sharman, J.L., Southan, C., Davies, J.A. The Concise Guide to PHARMACOLOGY 2015/16: G protein-coupled receptors. British Journal of Pharmacology 172:5744–5869 (2015)

[3] Albizu L., Cottet, M., Kralikova, M., Stoev, S., Seyer, R., Brabet, I., Roux, T., Bazin, H., Bourrier, E., Lamarque, L., et al. (2010) Time-resolved FRET between GPCR ligands reveals oligomers in native tissues. *Nat. Chem. Biol.* 6, 587-594.

[4] Altschul, S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389-3402.

[5] Andras P. (2002). Kernel-Kohonen networks. *International Journal of Neural Systems*, 12:117-135.

[6] Atkinson, H.J., Morris, J.H., Ferrin, T.E. and Babbitt, P.C.(2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One*, 4(2):e4345.

[7] Attwood, T.K. (2001) A Compendium of Specific Motifs for Diagnosing GPCR Subtypes. *Trends in Pharmacological Sciences*, 22: 162-165.

[8] Attwood, T.K. and Flower, D.R. (2002) Trawling the genome for G protein-coupled receptors: the importance of integrating bioinformatic approaches. *Special Publications of the Royal Society of Chemistry*, 279:60-71.

[9] Attwood, T.K., Bradley, P., Gaulton, A., Maudling, N., Mitchell, A.L. and Moulton, G. (2004) The PRINTS protein fingerprint database: functional and evolutionary applications. *In Encyclopaedia of Genetics, Genomics, Proteomics and Bioinformatics*, M.Dunn, L. Jorde, P.Little and A. Subramaniam (Eds.). John Wiley and Sons.

[10] Aupetit, M., Sedlmair, M. (2016) SepMe: 2002 new visual separation measures. *IEEE Pacific Visualization Symposium*, 1-8.

[11] Baldi, P. and Brunak, S.(2001) Bioinformatics: The Machine Learning Approach, MIT Press.

[12] Barnes P.J. (2006) Receptor heterodimerization a new level of cross-talk. *Journal of Clinical Investigation*, 116(5):1210-1212.

[13] Baum, D. (2008) Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups. *Nature Education.* 1(1):190

[14] Bediaga, N.G., Marichalar-Mendia, X., Aguirre-Urizar, J.M., Calvo, B., Echebarria-Goicouria, M.A., Pancorbo, M.M., Acha-Sagredo, A. (2014) Global DNA methylation: uncommon event in oral lichenoid disease. *Oral diseases*, 20(8): 821-826.

[15] Benz, V., Ghanshamdas, N. and Stevenson, J.P. (2015) Introducing Spain and its Pharmaceuticals Industry. *Industry Explorations. Spain Pharmaceuticals 2015. Global Business Reports.* 8-13

[16] Berry, M. and Linoff, G. (1997) Data Mining Techniques for Marketing, Sales, and Customer Support. John Wiley and Sons, Inc., New York.

[17] Bhasin M, Raghava GPS (2004) GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic Acids Res 32(suppl 2):W383-W389

[18] Bishop C.M. and Tipping M.E. (1998) A Hierarchical Latent Variable Model for Data Visualisation, IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 3(20):281-293.

[19] Bishop C.M., Svensén M. and Williams C. K. I. (1998) GTM: The Generative Topographic Mapping, *Neural Computation*, 10(1):215-234.

[20] Bishop C. M., Svensén M. and Williams C. K. I. (1998) Developments of the Generative Topographic Mapping, *Neurocomputing*, 21(1-3):203-224.

[21] Briesemeister, S., Rahnenführer, J. and Kohlbacher, O. (2012) No Longer Confidential: Estimating the Confidence of Individual Regression Predictions. *PLoS ONE 7(11): e48723.*

[22] Brown, B. (1996) Biological Membranes.*The Biochemical Society.*

[23] Bockaert, J. and Pin, J.P. (1999) Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO Journal*, 18(7):1723-1729.

[24] Bouvier, M. (2001) Oligomerization of G-protein-coupled transmitter receptors. *Nat. Rev. Neuroscience.* 2:274-286.

[25] Bruno, A., Constantino, G., de Fabritiis, G., Pastor, M., Selent, J. (2012) Membrane-sensitive conformational states of helix 8 in the metabotropic Glu2 receptor, a class C GPCR. *PloS ONE*, 7(8):e42023.

[26] Cano, G., García-Rodríguez, J., Orts-Escolano, S., Peña-Garcia, J., Kumar-Yadav, D., Pérez-Garrido, A., Pérez-Sánchez, H. (2015) Support Vector Machine prediction of drug solubility on GPUs. *Bioinformatics and Biomedical Engineering*, pp. 645-654, Springer.

[27] Cárdenas, M.I., Vellido, A. and Giraldo, J. (2014) Visual interpretation of class C GPCR subtype overlapping from the nonlinear mapping of transformed primary sequences. *In Proceedings of the International Conference on Biomedical and Health Informatics (IEEE BHI 2014)*, 764-767.

[28] Cárdenas, M.I., Vellido, A., König, C., Alquézar, R. and Giraldo, J. (2014) 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.

[29] Cárdenas, M.I., Vellido, A., König, C., Alquézar, R. and Giraldo, J. (2015). Visual characterization of misclassified class C GPCRs through manifold-based machine learning methods. *Genomics and Computational Biology*, 1(1):e19

[30] Christopoulos, A. and Kenakin, T. (2002) G Protein-Coupled Receptor Allosterism and Complexing. *Pharmacological Reviews*, 54:323-374,

[31] Cooke, R.M., Brown, A.J.H., Marshall, F.H., Mason, J.S. (2015) Structures of G protein-coupled receptors reveal new opportunities for drug discovery. *Drug Discovery Today*, 20:1355-1364.

[32] Cournia, Z., Allen, T.W., andricioaei, I., Anthony, B., Baum, D., Brannigan, G., Buchete, N.V., Deckman, J.T., Delemotte, L., del Val, C., Friedman, R. (2015) Membrane protein structure, function, and dynamics: a perspective from experiments and theory. *The Journal of Membrane Biology*, 248(4):611-640.

[33] Cruz-Barbosa, R., Vellido, A. and Giraldo, J. (2013) 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 2013)*, 759-766.

[34] Cruz-Barbosa, R., Vellido, A., Giraldo, J. (2015) The influence of alignment-free sequence representations on the semi-supervised classification of Class C G Protein-Coupled Receptors. Med Biol Eng Comput 53(2):137-149.

[35] Cuthbertson, J.M., Doyle, D.A., Sansom, M.S. (2005) Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Engineering, Design and Selection.* 18:295-308

[36] *http://www.cytoscape.org*

[37] Heterodimer. (n.d.) Dorland's Medical Dictionary for Health Consumers. (2007). Retrieved January 21 2012 from http://medical-dictionary.thefreedictionary.com/Heterodimer

[38] DasGupta, A. (2011) Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics. *Springer Texts in Statistics*

[39] Davies, M.N., Gloriam, D.E., Secker, A., Freitas, A.A., Timmis, J., Flower, D.R. (2011) Present perspectives on the automated classification of the G-protein coupled receptors (GPCRs) at the protein sequence level. Curr Top Med Chem 11(15):1994-2009

[40] Dempster, A. P., Laird N. M. and Rubin D. B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, B, 39(1):1-38.

[41] Dong Q., Zhou S., Guan J. (2009) A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics*, 25: 2655-2662.

[42] Dorè, A.S., Okrasa, K., Patel, J.C., Serrano-Vega, M., Bennett, K., Cooke, R.M., Errey, J.C., Jazayeri, A., Khan, S., Tehan, B., Weir, M., Wiggin, G.R. and Marshall, F.H. (2014) Structure of a class C GPCR metabotropic glutamate receptor 5 transmembrane domain. *Nature*. 551: 557-562.

[43] Durbin, R., Eddy S. R., Krogh A., and Mitchison G. (2004) Biological sequence analysis: Probabilistic models of proteins and nucleic acids. *Cambridge University Press*, Cambridge.

[44] Eddy, S.R. (2004) Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology*, 22(8):1035-1036.

[45] Elo, L. L. and Schwikowski, B. (2011) Mining proteomic data for biomedical research. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. doi: 10.1002/widm.45

[46] Eo, H.S. et al. (2009) A machine learning based method for the prediction of G protein-coupled receptor-binding PDZ domain proteins. *Molecules and Cells*, 27: 629-634.

[47] Faloutsos, Ch. and Lin, K., (1995) FastMap: a Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. *Proceedings of the ACM-SIGMOD international Conference on Management of Data*, 163-174

[48] Feenstra, K.A., Bastianelli, G., Heringa, J. (2008) Predicting protein interactions from funtional specificity. *From Computational Biophysics to Systems Biology* (CBSB08), 40, 89-92.

[49] Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods in Enzymology.*, 266:418-27.

[50] Feng, D.F., Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Molecular Biology and Evolution* 25:351-60.

[51] Filizola, M. and Weinstein, H. (2005)The study of G-protein coupled receptor oligomerization with computational modeling and bioinformatics. *FEBS Journal*, 272:2926-2938.

[52] Filmore, D. (2004) Cell-based screening assays and structural studies are fueling G-protein coupled receptors as one of the most popular classes of investigational drug targets. *Modern Drug Discovery*, 7(11).

[53] Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., Punta, M. (2014) The Pfam protein families database. *Nucleic Acids Research*, Database Issue 42:D222-D230.

[54] Fitch, W. M. (2000) Homology: a personal view on some of the problems. *Trends in Genetics*, 16(5):227-231.

[55] Foreman J. C. and Johansen T. (2003) Textbook of receptor pharmacology. CRC Press, 2nd edition.

[56] Fredriksson, R. et al. (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Molecular Pharmacology*, 63(6):1202-1205.

[57] Frénay, B., Verleysen, M. (2014) Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks.* 25(5):845-869

[58] Furukawa T.(2009) SOM of SOMs. *Neural Networks*, 22(4):463-478.

[59] Fuxe, K. et al. (2008) Heterodimers and Receptor Mosaics of Different Types of G-Protein-Coupled Receptors. *Physiology*, 23:322-332.

[60] Gama L., Wilt S.G., and Breitwieser G.E. (2001) Heterodimerization of calcium sensing receptors with metabotropic glutamate receptors in neurons. *The Journal of Biological Chemistry*, 276:39053-39059.

[61] Gantt, H. (1910) Work, Wages and Profit.*The Engineering Magazine.*

[62] Gao Q.B., Wang, Z.Z. (2006) Classification of G-protein coupled receptors at four levels. *Protein Eng Des Sel.* 19:511-516

[63] Gao, Q.B., Ye, X.F., He, J. (2013) Classifying G-Protein-Coupled Receptors to the Finest Subtype Level. *Biochem Bioph Res Co* 439(2), 303-308.

[64] García-Torres, M., Armañanzas, R., Bielza, C., and Larrañaga, P. (2013) Comparison of metaheuristic strategies for peakbin selection in proteomic mass spectrometry data. *Journal of Information Sciences* 222: 229-246.

[65] Gasparini F., Kuhn, R. and Pin, J. P. (2002) Allosteric Modulators of Group I Metabotropic Glutamate Receptors: Novel Subtype-Selective Ligands and Therapeutic Perspectives.*Current Opinion in Pharmacology*, 2:43-49.

[66] Gaulton, A. and Attwood, T.K.(2003) Bioinformatics approaches for the classification of G-protein-coupled receptors. *Current Opinion in Pharmacology*, 3:114-120.

[67] Geng, Y., Mosyak, L., Kurinov, I., Zuo, H., Sturchler, E., Cheng, T., et al. (2016). Structural mechanism of ligand activation in human calcium-sensing receptor. *Elife* 5:e13662. doi: 10.7554/eLife.13662

[68] Goudet, C., Binet, V., Prezeau, L. and Pin, J. P. (2004) Allosteric modulators of class-C G-protein-coupled receptors open new possibilities for therapeutic application. *Drud Discovery Today: Therapeutic Strategies*. 1(1): 125-133.

[69] Greene, D. et al (2008) Ensemble non-negative matrix factorization methods for clustering protein-protein interactions. *Bioinformatics*, 24(15):1722-1728.

[70] Guo, Y.Z., Li, M., Lu, M., Wen, Z., Wang, K., Li, G., Wu, J. (2006) Classifying G protein-coupled receptors and nuclear receptors on the basis of protein power spectrum from fast Fourier transform. *Amino Acids* 30(4):397-402.

[71] Haack, K.K.V., McCarty, N.A. (2011) Functional consequences of GPCR heterodimerization: GPCRs as allosteric modulators, *Pharmaceuticals*, 4(3):509-523.

[72] Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., Zhou, Y. (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports* 5:11476.

[73] Henikoff, J. G. and Henikoff, S. (1996) Using substitution probabilities to improve position-specific scoring matrices. *Computer Applications in the Biosciences*, 12: 135-143.

[74] Henikoff, S. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences (PNAS)*, 89:10915-10919

[75] Heskes, T. (1998). Energy functions for self-organizing maps. *Kohonen Maps*, 303-316.

[76] Higgs, P.G. and Attwood, T.K. (2005), *Bioinformatics and Molecular Evolution*. Blackwell Science.

[77] Horn F., Weare J., Beukers M. W., Horsch S., Bairoch A., Chen W., Edvardsen O., Campagne F. and Vriend G. (1998) GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acids Research*, 26:275-279.

[78] Hou Y., Hsu W., Lee M.L., Bystroff C. (2003) Efficient remote homology detection using local structure. Bioinformatics 19:2294-2301

[79] Hsu, A.L., Halgamuge, S.K. (2003) Enhancement of topology preservation and hierarchical dynamic self-organising maps for data visualization, *International Journal of Approximate Reasoning*. 32(2-3):259-279

[80] Hui, S., Xing, X., Bader, G.D. (2013) Predicting PDZ domain mediated protein interactions from structure. *BMC Bioinformatics* 14(1):27.

[81] Ideker, T., Ozier, O., Schwikowski, B. and Siegel, A.F.(2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 (Suppl 1):S233-S240.

[82] Iqbal, M.J., Faye, I., Samir, B.B. (2016) Classification of GPCRs proteins using a statistical encoding method. In International Joint Conference on Neural Networks (IJCNN) IEEE. 1224-1228.

[83] Isberg V., Mordalski, S., Munk, C., Rataj, K., Harpsoe, K., Hauser, A.S., Vroling, B., Bojarski, A.J., Vriend, G., Gloriam, D.E. (2016) GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acid Research*, 44(D1):D356-364.

[84] Jacoby, E., Bouhelal, R., Gerspacher, M. and Seuwen, K. (2006) The 7 TM G-protein coupled receptor target family. *ChemMedChem Chemistry Enabling Drug Discovery*, 1(8):761-782.

[85] Jain, K.K. (2004) Role of pharmacoproteomics in the development of personalized medicine. *Pharmacogenomics.* 5(3): 331-336.

[86] Jolliffe, I.T. (2002) Principal Component Analysis, *Springer Series in Statistics (2nd edition).* Springer, NY.

[87] Johansson, F. and Toh, H. (2010) A comparative study of conservation and variation scores. *Bioinformatics.* 11:388.

[88] Jain, K.K. (2004) Role of pharmacoproteomics in the development of personalized medicine. *Pharmacogenomics*, 5(3): 331-336.

[89] Jamali, A.A., Ferdousi, R., Razzaghi, S., Li, J., Safdari, R., Ebrahimie, E. (2016) DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. *Drug Discovery Today.* In Press.

[90] Karchin, R., Karplus, K., Haussler, D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* 18(1):147-159

[91] Karlin, S. and Rinott, Y. (1981) Entropy Inequalities for Classes of Probability Distributions I. The Univariate Case. *Advances in Applied Probability*, 13(1):93-112.

[92] Katoh. K., Misawa, K., Kuma, K. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acid Research* 30(14):3059-3066.

[93] Katoh, K., Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30:772-80.

[94] Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H.(2008) Visual analytics: scope and challenges. In Visual Data Mining: Theory, Techniques and Tools for Visual Analytics, LNCS 4404:77-90.

[95] Kersting, K. et al. (2010) Hierarchical Convex NMF for Clustering Massive Data. *Journal of Machine Learning Research*, 13:253-268.

[96] Khan, A., Khan, M.F., Choi, T.S. (2008) Proximity based GPCRs prediction in transform domain. Biochem Biophys Res Commun 371(3):411-415.

[97] Khelashvili, G., Dorff, K., Shan, J., Camacho-Artacho, M., Skrabanek, L., Vroling, B., Bouvier, M., Devi, L.A., George, S.R., Javitch, J.A., et al. (2010) GPCR-OKB: the G protein coupled receptor oligomer knowledge base. *Bioinformatics Appl. Note.* 26, 1804-1805.

[98] Kim, C. H., Lee, J., Lee, J. Y., and Roche, K. W. (2008) Metabotropic glutamate receptors: phosphorylation and receptor signaling. *Journal of Neuroscience Research* 86, 1-10.

[99] Kim, W. et al. (2011) Sparse nonnegative matrix factorization for protein sequence motif discovery. *Expert Systems with Applications*, 38:13198-13207.

[100] Klabunde, T. (2007) Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *British Journal of Pharmacology*, 152(1):5-7.

[101] Kobilka, B.K. (2012) The Structural Basis of G Protein Coupled Receptor Signaling. *The Nobel Prizes*, 195-213.

[102] Kohonen T. (1982) Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59-69.

[103] Kohonen T. (2001) Self-Organizing Maps (3rd ed). Springer-Verlag, Berlin.

[104] Kolakowski Jr L.F. (1994) GCRDb: a G-protein-coupled receptor database. *Receptor Channels.* 2(1):1-7.

[105] Kondor, R.I., and Lafferty, J.D.(2002) Diffusion kernels on graphs and other discrete input spaces. In Procs. of the ICML, San Francisco, CA, 315-322.

[106] König, C., Cruz-Barbosa, R., Alquézar, R. and Vellido, A. (2013) 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.

[107] König CL, Cárdenas MI, Giraldo J, Alquézar R, Vellido A (2015) Label noise in subtype discrimination of class C G-protein coupled receptors: A systematic approach to the analysis of classification errors. BMC Bioinf 16(1):314

[108] König C, Alquézar R, Vellido A, Giraldo J (2015) The extracellular N-terminal domain suffices to discriminate class C G Protein-Coupled Receptor subtypes from n-grams of their sequences. In Procs. of the International Joint-Conference on Artificial Neural Networks (IJCNN 2015), Killarney, Ireland, pp.2330-2336.

[109] Kroeger, K.M., Pfleger, K.D. and Eidne K.A. (2003) G-protein coupled receptor oligomerization in neuroendocrine pathways. *Frontiers in Neuroendocrinology*, 24:254-278.

[110] Krogh A., Larsson B., von Heijne, G. and Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a model: Application to complete genomes. *Journal of Molecular Biology*, 305:567-580.

[111] Kuang D., Yao Y., Wang M., Pattabiraman N., Kotra L.P., Hampson, D.R. (2003) Molecular similarities in the ligand binding pockets of an odor-

ant receptor and the metabotropic glutamate receptors. *Journal of Biological Chemistry*, 278(43): 42551-42559.

[112] Kunishima, N., Shimada, Y., Tsuji, Y., Sato, T., Yamamoto, M., Kumasaka, T., et al. (2000). Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor. *Nature* 407, 971-977.

[113] Lampinen J. and Oja E. (1992). Clustering properties of hierarchical self-organizing maps. *Journal of Mathematical Imaging and Vision*, 2:261-272.

[114] Lapinsh, M., Gutcaits, A., Prusis, P., Post, C., Lundstedt, T., Wikberg, J.E.S.(2002) Classification of G-protein Coupled Receptors by Alignment-Independent Extraction of Principal Chemical Properties of Primary Amino Acid Sequences. Protein Sci. 11(4):795-805.

[115] Lau K.W., Yin H. and Hubbard S. (2006) Kernel self-organising maps for classification. *Neurocomputing*, 69:2033-2040.

[116] Lee, DD. and Seung, HS. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788-791.

[117] Lee, J.A., Verleysen, M.(2007) Nonlinear Dimensionality Reduction. Springer.

[118] Lefkowitz, R.J. (2013) A Brief History of G-Protein Coupled Receptors (Nobel Lecture). *Angewandte Chemie International Edition*, 52, 2-15.

[119] Li, Y. et al. (2012) Hierarchical non-negative matrix factorization (hNMF): a tissue pattern differentiation method for glioblastoma multiforme diagnosis using MRSI. *NMR in Biomedicine.*

[120] Libbrecht, M.W., Noble, W.S. (2015) Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321-332.

[121] Lipman D.J.and Pearson W.R. (1985). Rapid and sensitive protein similarity searches. *Science* 227(4693):1435-1441

[122] Lisboa, P.J.G., Vellido, A., Tagliaferri,R., Napolitano, F., Ceccarelli, M., Martin-Guerrero, J.D. and Biganzoli, E. (2010) Data Mining in Cancer Research, *IEEE Computational Intelligence Magazine*, 5(1): 14-18.

[123] Liu X., Zhao L., Dong Q. (2011) Protein remotehomologydetectionbasedonauto-cross covariance transformation. *Computers in Biology and Medicine*, 41: 640-647.

[124] Lohse, M. (2010) Dimerization in GPCR mobility and signaling. *Curr. Opin. Pharmacol.* 10, 53-58.

[125] MacDonald D. and Fyfe C.(2000) The Kernel Self Organising Map. *Applied Computational Intelligence Research Unit*, The University of Paisley, Scotland.

[126] Mamoshina, P., Vieira, A., Putin, E., Zhavoronkov, A. (2016) Applications of deep learning in biomedicine. *Molecular Pharmaceutics.* In Press.

[127] Maniyar D.M. (2006) Data Visualization during the Early Stages of Drug Discovery. *Journal of Chemical Information and Modelling* , 46:1806-1818.

[128] Márquez-Chamorro, A.E., Asencio-Cortés, G., Santiesteban-Toca, C.E., Aguilar-Ruiz, J.S. (2015) Soft computing methods for the prediction of protein tertiary structures: A survey. *Applied Soft Computing*, 35: 398-410.

[129] Masegosa A., Armañanza R., Grau M.A., Potenciano V., Moral S., Larrañaga P., Bielza C., Matesanz F. (2015) Discretization of expression quantitative trait loci in association analysis between genotypes and expression data. *Current Bioinformatics* 10(2): 144-164.

[130] Mendelson, S. and Neeman, J. (2010) Regularization in kernel learning. *Institute of Mathematical Statistics. The annals of statistics*, 38(1):526-565.

[131] Meila, M. (2007) Comparing clusterings an information based distance. *Journal of Multivariate Analysis*, 98(5):873-895.

[132] Milligan, G. (2009) G protein-coupled receptor hetero-dimerization: contribution to pharmacology and function. *Br. J. Pharmacology.* 158, 5-14.

[133] Milligan, G. (2013) The prevalence, maintenance, and relevance of G protein-coupled receptor oligomerization. *Molecular Pharmacology*, 84(1):158-169.

[134] Mirzadegan, T.,Benkö, G., Filipek, S. and Palczewski, K.(2003) Sequence analyses of G protein-coupled receptors: similarities to Rhodopsin. *Biochemistry*, 42(10):2759-2767.

[135] Mohabatkar, H., Beigi, M.M., Esmaeili, A. (2011) Prediction of GABA A receptor proteins using the concept of Chou's pseudo-amino acis composition and support vector machine. *Journal of the theoretical Biology* 281(1):18-23.

[136] Moller, S., Croning, M.D., Apweiler, R. (2001) Evaluation of Methods for the Prediction of Membrane Spanning Regions. *Bioinformatics*, 17:646-653.

[137] Moya (2008) Informe Moya-Angeler, directed by Joaquín Moya-Angeler for Farmaindustria, October 2008.

[138] Munk, C., Isberg, V., Mordalski, S., Harpsøe, K., Rataj, K., Hauser, A. S., Kolb, P., Bojarski, A. J., Vriend, G. , and Gloriam, D. E. (2016) GPCRdb: the G protein-coupled receptor database an introduction. *Br J Pharmacol*, May 8. 10.1111/bph.13509

[139] Muto, T., Tsuchiya, D., Morikawa, K., and Jingami, H. (2007). Structures of the extracellular regions of the group II/III metabotropic glutamate receptors. *Proc. Natl. Acad. Sci.* U.S.A. 104, 3759-3764.

[140] Nabney, I.T., Sun, Y., Tiňo, P. and Kabán, A. (2005) Semisupervised Learning of Hierarchical Latent Trait Models for Data Visualization. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):384-399.

[141] Nada, L., Elpida, K. and Blaž, Z. (2012) Intelligent Data Analysis in Medicine and Pharmacology. Springer Science & Business Media. ISBN 978-1-4613-7775-7.

[142] Nisius, B. and Bajorath, J. (2011) Mapping of pharmacological space. *Expert Opinion Drug Discovery*, 6(1).

[143] Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A novel method for multiple sequence alignments. *Journal of Molecular Biology*, 302:205-217.

[144] Oakes, M. (1998) Statistics for Corpus Linguistics. *International Journal of Applied Linguistics*, 10(2):269-274.

[145] Ogul, H., Mumcuoglu, E.U. (2007) A discriminative method for remote homology detection based on n-peptide compositions with reduced amino acid alphabets. BioSystems 87:75-81.

[146] Olier, I., Vellido, A. and Giraldo, J. (2010) Kernel Generative Topographic Mapping. *ESANN 2010 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 481-486.

[147] Oliveira, L., Paiva, P.B., Paiva, A. and Vriend, G. (2003) Sequence analysis reveals how G protein-coupled receptors transduce the signal to the G protein. *PROTEINS: Structure, Function and Genetics*, 52:553-560.

[148] Otaki, J.M., Mori, A., Itoh, Y., Nakayama, T., Yamamoto, H. (2006) Alignment-free classification of G-protein-coupled receptors using self-organizing maps. J Chem Inf Model 46(3):1479-1490

[149] Paatero, P. and Tapper, U. (1994) Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5: 111-126.

[150] Paliwal, K., Lyons, J., Heffernan, R. (2015) A short review of Deep Learning Neural Networks in protein structure prediction problems. *Advanced Techniques in Biology & Medicine*, 3:3.

[151] Page R.D.M. and E.C. Holmes. (1998) Molecular Evolution: A Phylogenetic Approach. Blackwell Science. Oxford. Chapter 2. p.11.

[152] Pais, F.S., Ruy, PdC., Oliveira, G. and Coimbra, R.S. (2014) Assessing the efficiency of multiple sequence alignment programs. *Algorithms for Molecular Biology*, 9:4.

[153] Pan, Z. S., Chen, S. C. and Zhang, D. Q. (2004). A kernel-base SOM classifier in input space. *Acta Electronica Sinica*, 32: 227-231.

[154] Pearson, K. (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(6):559-572.

[155] Peng, Z.L., Yang, J.Y., Chen, X. (2010) An improved classification of G-protein-coupled receptors using sequence-derived features. *BMC Bioinf.* 11:420

[156] Pin, J-P. and Duvoisin, R. (1995) The metabotropic glutamate receptors: structure and functions. *Neuropharmacology* 34: 1-26.

[157] Pin, J-P., Galvez T., Prezeau, L.(2003) Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. *Pharmacology and Therapeutics*, 98:325-354.

[158] Pharmaceutical Industry Competitiveness Task Force (PICTF), Department of Health, U.K. URL: http://www.dh.gov.uk/ab/Archive/PICTF/index.htm

[159] Pisier, G. (1989). The Volume of Convex Bodies and Banach Space Geometry. Cambridge Univ. Press, Cambridge.

[160] Pittolo S. et al. (2014) An allosteric modulator to control endogenous G protein-coupled receptors with light. *Nature Chemical Biology*, 10: 813-815

[161] Prinster, S.C., Hague C. and Hall, R.A.(2005) Heterodimerization of G Protein-Coupled Receptors: Specificity and Functional Significance. *Pharmacological Reviews*, 57(3):289-298.

[162] Qi, Y., Oja, M., Weston, J., Noble, W.S. (2012) A unified multitask architecture for predicting local protein properties. *PLoS ONE* 7(3):e32235.

[163] Qi, Y., Das, S.G., Collobert, R., Weston, J. (2014) Deep learning for character-based information extraction. In *Advances in Information Retrieval*, pp.668-674, Springer.

[164] Rabiner L.R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77(2).

[165] Rauber, A. et al. (2002) The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. *IEEE Transactions on Neural Networks*, 13(6).

[166] Rehman, Z.U., Mirza, M.T., Khan, A., Xhaard, H. (2013) Predicting g-protein-coupled receptors families using different physiochemical properties and pseudo amino acid composition. Method enzymol 522, 61-79.

[167] Rick, N.G. (2015) Drugs, from discovery to approval. John Wiley & Sons, Inc. ISBN: 978-1-118-90727-6

[168] Rozenfeld, R., Devi, L. (2010) Receptor heteromerization and drug Discovery. *Trends Pharmacol. Sci.* 31, 124-130.

[169] Saitou N., Nei M. (1987) The Neighbor-Joining Method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406-425.

[170] Sánchez, P.L., Las grandes cifras de la industria y del mercado farmacéutico español; análisis económico de la I+D farmacéutica. Farmaindustria, Barcelona, 28 de noviembre de 2008.

[171] Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M. and Wold, S.(1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *Journal of Medicinal Chemistry* 41:2481-2491.

[172] Sanders, M. et al. (2011) ss-TEA: Entropy based identification of receptor specific ligand binding residues from a multiple sequence alignment of class A GPCRs. *BMC Bioinformatics*, 12:332.

[173] San Miguel Hernández, Á. (2011) Importancia de la implantación de la proteómica a nivel asistencial. Revista del Laboratorio Clínico, 4(4):171-172.

[174] Santamaría, R., Therón, R. (2009) Treevolution: visual analysis of phylogenetic trees. *Bioinformatics*, 25(15), 1970-1971

[175] Schölkopf, B., Smola, A. and Müller K.R. (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299-1319.

[176] Schölkopf B. and Smola A. (2002) Learning with Kernels. The MIT Press, Cambridge, Massachusetts.

[177] Shah, A.R., Oehmen, C.S., Webb-Robertson, B.J. (2008) SVM-HUSTLE - an iterative semi-supervised machine learning approach for pairwise protein remote homology detection. *Bioinformatics* 24(6):783-790.

[178] Shannon, P. et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498-2504.

[179] Shawe-Taylor, J. and Cristianini, N. (2004) Kernel Methods for Pattern Analysis. *Cambridge University Press.*

[180] Sips, M., Neubert, B., Lewis, J.P., Hanrahan, P. (2009) Selecting good views of high-dimensional data using class consistency. *Comput Graphics Forum.* (28)3:831-838.

[181] Sievers, F. et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539.

[182] Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195-197.

[183] Sokal R. and Michener C. (1958) A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38: 1409-1438.

[184] Sonnhammer, E.L.L., von Heijne, G., and Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *In Procedure of Sixth International Conference on Intelligent Systems for Molecular Biology*, 175-182, AAAI Press.

[185] Srivastava, P.K., Desai, D.K., Nandi, S., Lynn, A.M. (2007) HMM-ModE-Improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. BMC Bioinf 8(1):1.

[186] Strope, P.K., Moriyama, E.N. (2007) Simple alignment-free methods for protein classification: a case study from G-protein-coupled receptors. *Genomics* 89(5):602-612.

[187] Takeda, Sh. et al. (2002) Identification of G protein-coupled receptor genes from the human genome sequence. *FEBS Letters*, 520:97-101.

[188] Tan, P., Steinbach, M. and Kumar, V. (2006) Introduction to Data Mining. *Addison Wesley.*

[189] Tautermann, C.S. (2014) GPCR structures in drug design, emerging opportunities with new structures. *Bioorg Med Chem Lett*, 24:4073-4079.

[190] Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673-4680.

[191] Thurau, C. et al. (2009) Convex non-negative matrix factorization in the wild. In H. Kargupta and W. Wang, editors, *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM-09).*

[192] Tiňo, P. and Nabney, I. (2002) Hierarchical GTM: constructing localized non-linear projection manifolds in a principled way. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):639-656.

[193] Tsuchiya, D., Kunishima, N., Kamiya, N., Jingami, H., and Morikawa, K. (2002). Structural views of the ligand-binding cores of a metabotropic glutamate receptor complexed with an antagonist and both glutamate and Gd3+. *Proc. Natl. Acad. Sci.* U.S.A. 99, 2660-2665. doi: 10.1073/pnas.052708599

[194] Ultsch, A., Mörchen, F. (2005) ESOM-Maps: Tools for clustering, visualization, and classification with Emergent SOM. Technical Report 46, CS Department, Philipps-University Marburg, Germany.

[195] Vellido, A., Lisboa, P.J.G. and Meehan, K. (2000) The generative topographic mapping as a principal model for data visualization and market

segmentation: an electronic commerce case study. *International Journal of Computer Systems and Signals*, 1(2):119-138.

[196] Vellido, A., Martín, J.D., Rossi, F., and Lisboa, P.J.G. (2011) Seeing is believing: The importance of visualization in real-world machine learning applications,*In Procs. of the 19th European Symposium on Artificial Neural Networks (ESANN 2011)*, 219-226.

[197] Vellido A., Cárdenas M.I., Olier I., Rovira X. and Giraldo J. (2011) A probabilistic approach to the visual exploration of G Protein-Coupled Receptor sequences. In Procs. of the 19th European Symposium on Artificial Neural Networks (ESANN 2011), 233-238.

[198] Vellido, A., Martín-Guerrero, J.D., Lisboa, P.J.G.(2012) Making machine learning models interpretable, *in Proc. 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges*. 163-172.

[199] Vroling, B., Sanders, M., Baakman, C., Borrmann, A., Verhoeven, S., Klomp, J., Oliveira, L., de Vlieg, J., Vriend, G.(2011) GPCRdb: information system for G proteincoupled receptors. *Nucleic Acids Res 39(suppl 1)*, D309-D319

[200] Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. and Barton, G.L. (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189-1191.

[201] Webb-Robertson, B.J., Oehmen, C., Matzke, M. (2005) SVM-BALSA: Remote homology detection based on Bayesian sequence alignment. Comput Biol Chem 29(6):440-443.

[202] Weill, N. and Rognan, D. (2009) Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G

protein-coupled receptors and their ligands. *Journal of Chemical Information and Modeling*, 49:1049-1062.

[203] Wistrand, M., Käll, L. and Sonnhammer, E.L.L. (2005) A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Protein Science*, 15:509-521.

[204] Williams, C.K.I. (2000) A MCMC Approach to Hierarchical Mixture Modelling. *Advances in Neural Information Processing Systems 12*. 680-686.

[205] Wu, H., Wang, K.J., Han, G.W., Cho, K.P., Xia, C.M., Katritch, J., Meiler, J., Cherezov, P., Conn, J. and Steven, R.C. (2014) Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. *Science*. 344(6179): 58-64.

[206] Xiao, X., Lin, W.Z., Chou, K.C. (2012) Recent advances in predicting G-protein Coupled Receptor classification. *Current Bioinformatics* 7(2):132-142.

[207] Yabuki, Y. et al.(2005) GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model. *Nucleic Acids Research*, 33:W148-W153.

[208] Ye, K. et al. (2006) A two-entropies analysis to identify functional positions in the transmembrane region of class A G-protein coupled receptors. *PROTEINS: Structure, Function and Bioinformatics*, 63:1018-1030.

[209] Ye, K., Feenstra, K.A., Heringa, J., Ijzerman, A.P. and Marchiori, E. (2008) Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. *Bioinformatics*, 24(1): 18-25.

[210] Ye, J. et al. (2008) Multi-class discriminant kernel learning via convex programming. *Journal of Machine Learning Research*, 9:719-758.

[211] Yin H. (2008) The Self-Organizing Maps: Background, Theories, Extensions and Applications, *Studies in Computational Intelligence* (SCI) 115:715-762.

[212] Yu, L.R. (2011) Pharmacoproteomics and toxicoproteomics: The field of dreams. *Journal of Proteomics*, 74:2549-2553.

[213] Zhang, C., Zhang, T., Zou, J., Miller, C. L., Gorkhali, R., Yang, J. Y., et al. (2016). Structural basis for regulation of human calcium-sensing receptor by magnesium ions and an unexpected tryptophan derivative co-agonist. *Sci Adv.* 2:e1600241. doi: 10.1126/sciadv.1600241