

**Analysis of class C G-Protein Coupled Receptors
using supervised classification methods**



Caroline Leonore König

Supervised by:

Dr. René Alquézar Mancho and Dr. Alfredo Vellido Alcacena

Computer Science Department

Universitat Politècnica de Catalunya

A thesis submitted for the degree of *Ph.D. in Artificial Intelligence*

Acknowledgments

I would like to thank my advisors Dr. Alfredo Vellido and Dr. René Alquézar from the SOCO research group of the UPC, for giving me the opportunity to develop my Ph.D. thesis as part of the KAPPA-AIM¹ project and work on the investigation of G protein-coupled receptors with Artificial Intelligence methods. Both of them helped me with questions and provided me useful feedback, as well as valuable advice and input at every stage of this thesis, spending a long time during the preparation of this work. As well, I would like to thank to Dr. Jesús Giraldo from the 'Institut de Neurociències' of the 'Universitat Autònoma de Barcelona' (UAB) for the large collaboration in this PhD research providing so many biological insight to the study.

¹KAPPA-AIM: Knowledge Acquisition in Pharmacoproteomics using Advanced Artificial Intelligence Methods

Abstract

G protein-coupled receptors (GPCRs) are cell membrane proteins with a key role in regulating the function of cells. This is the result of their ability to transmit extracellular signals, which makes them relevant for pharmacology and has led, over the last decade, to active research in the field of proteomics. The current thesis specifically targets class C of GPCRs, which are relevant in therapies for various central nervous system disorders, such as Alzheimer's disease, anxiety, Parkinson's disease and schizophrenia. The investigation of protein functionality often relies on the knowledge of crystal three dimensional (3-D) structures, which determine the receptor's ability for ligand binding responsible for the activation of certain functionalities in the protein. The structural information is therefore paramount, but it is not always known or easily unravelled, which is the case of eukaryotic cell membrane proteins such as GPCRs. In the face of the lack of information about the 3-D structure, research is often bound to the analysis of the primary amino acid sequences of the proteins, which are commonly known and available from curated databases. Much research on sequence analysis has focused on the quantitative analysis of their aligned versions, although, recently, alternative approaches using machine learning techniques for the analysis of alignment-free sequences have been proposed. In this thesis, we focus on the differentiation of class C GPCRs into functional and structural related subgroups based on the alignment-free analysis of their sequences using supervised classification models. In the first part of the thesis, the main topic is the construction of supervised classification models for unaligned protein sequences based on physicochemical transformations and n -gram representations of their amino acid sequences. These models are useful to assess the internal data quality of the externally labeled dataset and to manage the label noise problem from a data curation perspective. In its second part, the thesis focuses on the analysis of the sequences to discover subtype- and region-specific sequence motifs. For that, we carry out a systematic analysis of the topological sequence segments with supervised classification models and evaluate the subtype discrimination capability of each region. In addition, we apply different types of feature selection techniques to the n -gram representation of the amino acid sequence segments to find subtype and region specific motifs. Finally, we compare the findings of this motif search with the partially known 3D crystallographic structures of class C GPCRs.

Contents

List of abbreviations	xi
1. Introduction	1
1.1. Motivation	1
1.2. Objectives	2
1.3. Methodology	2
1.4. Main thesis contributions	3
1.4.1. List of publications	5
1.5. Organization of the thesis	5
2. Biological Background	7
2.1. Introduction	7
2.2. Proteomics	7
2.2.1. Protein structures	7
2.2.2. Sequence motifs	9
2.3. G Protein-Coupled Receptors	10
2.4. GPCRdb	13
2.4.1. Class C GPCR datasets	14
3. Technical background	19
3.1. Introduction	19
3.2. Supervised Learning Techniques	19
3.2.1. Classification Models	20
3.2.2. Metrics	23
3.2.3. Cross Validation	25
3.3. Alignment-free data representation	25
3.3.1. Transformations based on the physicochemical properties	25
3.3.2. N-gram representations	27
3.4. Feature Selection	29
3.4.1. t-test Filtering	30
3.4.2. <i>Chi - square</i> Filtering	31
3.4.3. Sequential Forward Selection	31
3.5. Analysis of Label Noise	31
3.5.1. Label noise detection	32

4. Construction of supervised classification models using alignment-free transformations	33
4.1. Experiments with the 2011 dataset	33
4.1.1. Experimental settings	33
4.1.2. Experiments with transformations based on the physicochemical properties	34
4.1.3. Experiments with n-gram transformations	36
4.1.4. Conclusion	39
4.2. Experiments with the September 2016 dataset	39
4.2.1. Experimental settings	39
4.2.2. Results and Discussion	40
4.2.3. Conclusion	41
5. A systematic approach to GPCR misclassification analysis	43
5.1. Introduction	43
5.2. Proof of concept for a novel systematic approach to assist the discovery of GPCR database labelling quality problems	44
5.2.1. A systematic approach to misclassification analysis	45
5.3. Experiments on the class C GPCR dataset	48
5.3.1. Experimental Settings	49
5.3.2. Mislabeling analysis of the 2011 dataset	49
5.3.3. Mislabeling analysis of the 2016 dataset	57
6. Tracking the evolution of class C GPCR database for biocuration assistance	61
6.1. Introduction	61
6.2. Comparison of class C GPCR datasets	62
6.2.1. Experimental settings	62
6.2.2. Experimental results	63
7. Topological sequence segments discriminate between class C GPCR subtypes	73
7.1. Introduction	73
7.2. Experiments on the class C GPCR dataset	74
7.2.1. Experimental Settings	74
7.2.2. Analysis of the discrimination capability of segments from the 2011 dataset	76
7.2.3. Analysis of the discrimination capability of segments from the September 2016 dataset	84

8. Discovering class C GPCR motifs	91
8.1. Introduction	91
8.2. Feature selection used for the identification of subtype-discriminating n-grams	92
8.2.1. Experiments on the complete sequences	92
8.2.2. Experiments on the N-terminus	103
8.3. Feature selection used for the identification of subtype characteristic n-grams	112
8.3.1. Experiments on the N-terminus with a two-stage Feature Selection	113
8.3.2. Experiments on the N-terminus using chi-square Filtering . .	122
9. Analysis of 3-D crystal structures	131
9.1. Introduction	131
9.2. Experiments with crystal structures of the N-terminus	131
9.2.1. Experimental settings	132
9.2.2. Results	136
9.2.3. Discussion	146
9.2.4. Conclusion	147
10. Conclusions and future work	159
Bibliography	161
A. Appendix	173
A.1. Figures of the misclassification analysis	173
A.2. List of frequent misclassified sequences	179

List of Figures

2.2.1. Protein structure levels	8
2.3.1. GPCR Signalling Pathways	10
2.3.2. Schematic representation of class C GPCRs	11
2.3.3. Orthosteric and allosteric binding sites.	12
2.4.1. Subtype distribution for the different datasets.	16
2.4.2. Phylogenetic tree representation of the 2011 dataset.	17
6.2.1. Subtype distribution for the different datasets (with orphans).	66
7.2.1. Transmembrane GPCR structure	75
8.2.1. Graphical representation of the n -gram frequencies (CSL, FSML and ITFS).	99
8.2.2. Box plot of the CSL n -gram	101
8.2.3. Box plot of the ITF n -gram	102
8.2.4. Box plot of the FSM n -gram	102
8.2.5. Frequencies boxplots of n -grams of the AA alphabet.	110
8.2.6. Frequencies boxplots of n -grams of the SEZ alphabet.	111
8.3.1. Position-weighted selection of n -grams for mG and CS.	117
8.3.2. Position-weighted selection of n -grams for GB and Ta.	117
9.2.1. Protein Data Bank Screen for structure 2E4Z.	134
9.2.2. UniprotKB entry for protein P23385.	135
9.2.3. Visualization of structure 4MQE	137
9.2.4. Example of surface rendering	138
A.1.1. Boxplot representation of the accuracy of the AA, Digram, ACC and PDBT dataset.	173
A.1.2. Boxplot representation of the MCC of the AA, Digram, ACC and PDBT dataset.	174
A.1.3. Phylogenetic tree representation of the 2011 version dataset.	175
A.1.4. Mislabelings predicted to be mG.	176
A.1.5. Mislabelings predicted to be Od.	177
A.1.6. Mislabelings predicted to be Ph.	178
A.1.7. Mislabelings predicted to be Vn.	179

List of Tables

1.1. Overview of publications	5
2.1. Overview of number of receptors per dataset version.	15
3.1. Performance measures for binary classifiers.	24
3.2. Performance measures for multi-class classifiers.	24
3.3. Amino acid alphabet.	27
3.4. Amino acid grouping schemes.	28
3.5. Size of n -gram feature spaces.	29
4.3. Subtype classification results for ACC (2011 dataset).	35
4.1. Comparison of classifiers (2011 dataset).	35
4.2. Model selection results (MEAN, ACC and PDBT).	35
4.4. Classification results for n -gram representations (2011 dataset).	37
4.5. Classification results for the Prot2Vec transformation (2011 dataset).	37
4.7. Subtype classification results N -gram representation (2011 dataset).	38
4.6. Model selection result for n -gram and Prot2Vec transformations.	38
4.8. Comparison of classifiers (2016 dataset).	40
4.9. Model selection result (2016 dataset).	40
4.10. Subtype classification results for ACC (2016 dataset).	41
5.1. Overview SVM classifier results (2011 dataset).	50
5.2. Illustrative example of misclassification statistics.	51
5.3. Sequences with large classification errors.	52
5.4. Overview SVM classifier results (2016 dataset).	58
5.5. Analysis of misclassification of sequences (2016 dataset).	58
6.1. Comparison of number of receptors by dataset version.	65
6.2. Comparison of classifiers (2011 dataset).	68
6.3. Comparison of classifiers (2016 datasets).	69
6.4. Subtype classification results for Digram (2011 dataset).	69
6.5. Subtype classification results for ACC (2016 datasets).	70

7.1. Number of sequences with 7TM structure per subtype (2011 dataset).	76
7.2. Statistical information segment lengths (2011 dataset).	77
7.3. Classification results complete sequences (2011 dataset).	78
7.4. Subtype classification results for Digram (2011 dataset).	78
7.5. Classification results extracellular segments (2011 dataset).	79
7.6. Classification results transmembrane segments (2011 dataset).	80
7.7. Classification results intracellular segments (2011 dataset).	80
7.8. Classification results N-terminus with 7TM regions (2011 dataset).	80
7.9. Classification results all 15 segments (2011 dataset).	81
7.10. Subtype classification results by sequence segments (2011 dataset).	82
7.11. Number of sequences with 7TM structure per subtype (2016 dataset).	84
7.12. Statistical information segment lengths (2016 dataset).	85
7.13. Classification results complete sequence (dataset 2016).	86
7.14. Subtype classification results for Digram (2016 dataset).	86
7.15. Classification results extracellular segments (2016 dataset).	87
7.16. Classification results transmembrane segments (2016 dataset).	88
7.17. Classification results intracellular segments (2016 dataset).	88
7.18. Classification results N-terminus with 7TM regions (2016 dataset).	88
7.19. Classification results all 15 segments (2016 dataset).	89
8.1. <i>N</i> -gram classification results without FS.	94
8.2. <i>N</i> -gram classification results using sequential forward feature selection.	95
8.3. Classification results with t-test-based subset selection.	96
8.4. Classification results with two stage feature selection.	96
8.5. Subtype classification result for the reduced SEZ dataset.	98
8.6. Classification comparison N-terminus and complete sequence.	105
8.7. <i>N</i> -gram comparative classification results after t-test.	106
8.8. Classification results by alphabets using sequential forward selection.	106
8.9. Lists of <i>n</i> -grams ranked by relevance (AA alphabet).	108
8.10. Lists of <i>n</i> -grams ranked by relevance (SEZ alphabet).	109
8.11. Subtype classification results after t-test (subtype- <i>vs</i> -all).	115
8.12. Subtype classification results after Forward Selection (subtype- <i>vs</i> -all).	115
8.13. Subclass classification results (subtype- <i>vs</i> -subtype).	116
8.14. Selected <i>n</i> -grams for each subtype.	118
8.15. Location of <i>n</i> -grams.	119
8.16. Size of <i>n</i> -gram feature space by dataset.	123
8.17. List of <i>n</i> -grams by χ^2 selection for Mg (2011 dataset).	124
8.18. List of <i>n</i> -grams by χ^2 selection for Mg (2016 dataset).	124

8.19. List of n -grams by χ^2 selection for CS (2011 dataset).	125
8.20. List of n -grams by χ^2 selection for CS (2016 dataset).	125
8.21. List of n -grams by χ^2 selection for GB (2011 dataset).	126
8.22. List of n -grams by χ^2 selection for GB (2016 dataset).	126
8.23. List of n -grams by χ^2 selection for Ta (2011 dataset).	127
8.24. List of n -grams by χ^2 selection for Ta (2016 dataset).	127
8.25. Comparison of classification results by feature selection approach.	128
9.9. Crystal structures for mG receptor subtype 2.	140
9.10. Detailed analysis of Q14416.	141
9.11. Crystal structures for mG receptor subtype 3.	142
9.12. Detailed analysis of P31422.	142
9.13. Crystal structures for mG receptor subtype 3 (Q14832).	143
9.14. Detailed analysis of Q14832.	143
9.15. Crystal structures for mG receptor subtype 5 (P41594).	144
9.16. Detailed analysis of P41594.	144
9.17. Crystal structures for mG receptor subtype 7 (P35400).	145
9.18. Detailed analysis of P35400	145
9.19. Crystal structures for mG receptor subtype 7 (Q14831).	146
9.1. Relation of crystal structures of mG extracellular domain.	149
9.2. Matching of subtype specific mG n -grams with crystal structures.	150
9.3. Relation of crystal structures of CS extracellular domain.	151
9.4. Matching of subtype specific CS n -grams with crystal structures.	151
9.5. Relation of crystal structures of GB extracellular domain.	152
9.6. Matching of subtype specific GB n -grams with crystal structures.	153
9.7. Crystal structures for mG receptor subtype 1.	154
9.8. Detailed analysis of P23385.	154
9.20. Detailed analysis of Q14831	155
9.21. Crystal structures for CS extracellular domain.	155
9.22. Detailed analysis of P41180.	156
9.23. Crystal structures for GB extracellular domain.	156
9.24. Detailed analysis of Q9UBS5.	157
A.1. Frequently misclassified sequences for subtypes mG, Cs, GB and VN.	180
A.2. Frequently misclassified sequences for subtypes Ph and Od.	181

List of abbreviations

γ	Gamma parameter of radial basis function
θ_R	Threshold of the voting ratio
θ_{CDV}	Threshold of the cumulative decision value
3-D	Three dimensional
5-CV	5-fold cross validation
7TM	Seven transmembrane helices
AA	Amino acid
AC	Auto covariance
ACC	Auto-Cross Covariance
Accu	Accuracy
C	Error penalty parameter
CBOW	Continuous bag-of-words
CC	Cross covariance
CDV_s	Cummulative decision value for sequence s
CRD	Cystein rich domain
CS	Calcium Sensing
CSG	Continuous skip-gram
DAV	Davies Random
DM	Data mining

DSB	Disulfide bond
DT	Decision tree
e	Misclassification threshold
EL	Extracellular loops
ELM	Eukaryotic Linear Motif
ER _s	Error rate for sequence s
fn	False negatives
fp	False positives
FS	Feature selection
GB	GABAB
GPCR	G-protein coupled receptor
GPCRdb	GPCR database
IL	Intracellular loops
IUPHAR	International Union of Pharmacology
LN	Label noise
MCC	Matthews correlation coefficient
MEME	Multiple EM for Motif Elicitation
mG	Metabotropic Glutamate
ML	Machine learning
MLP	Multi-Layer Perceptron
MnM	Mini Motif Miner
MSA	Multiple Sequence Alignment
NB	Naive Bayes
nc	Number of instances of the subtype

NLP	Natural language processing
NMR	Nuclear magnetic resonance
Od	Odorant
PDB	Protein data bank
PDBT	Physicochemical distance-based transformation
Ph	Pheromone
PT	Phylogenetic tree
R_s	Voting ratio for sequence s
RBF	Radial basis function
RF	Random forest
SEZ	Sezerman
SFFS	Sequential forward feature selection
SVM	Support Vector Machine
Ta	Taste
tc	Number of instances of the subtype matching the n-gram
tn	True negatives
to	Number of instances of the rest of subtypes matching the n-gram
tp	True positives
VFT	Venus Flytrap
VN	Vomeronasal
VP_s	Votes obtained by the most frequently predicted label class
VT_s	Votes obtained by the true label class

1. Introduction

1.1. Motivation

G protein-coupled receptors (GPCRs) are a large and heterogeneous superfamily of receptors that are key cell players for their role as extracellular signal transmitters. Class C GPCRs, in particular, are of great interest in pharmacology.

The functionality of a protein depends on its three dimensional (3-D) structure to a large extent, which determines its ability for certain ligand binding. The information about the 3-D structure is therefore paramount as the ligand binding process activates certain functionality in the protein. Nevertheless, the 3-D structure of a protein is commonly unknown, as this information is obtained by crystallography, which has a high cost and is extremely difficult in some cases. The latter is the case for GPCR proteins, as they are transmembrane proteins, which are not solvable. For this reason, only the 12% of the 3-D structure of the human GPCR superfamily [1] were known at the beginning of this investigation. In the face of the lack of information about the 3-D structure, research in proteomics must analyze the primary amino acid sequences of the proteins, which are commonly known and available from publicly accesible curated databases. This approach allows to infer information about the functionality of proteins with unknown 3-D structure by searching for similarities between sequences with known functionality, a procedure known as homology detection.

In the past, much research on protein sequence analysis has focused on the quantitative analysis of their aligned versions, although, recently, alternative approaches for the analysis of alignment-free sequences have been proposed from the field of machine learning (ML). In this thesis, we focus on the alignment-free analysis of proteins for homology detection. In particular, we work with the class C GPCRs, which have become an important research target for new therapies for pain, anxiety and neurodegenerative disorders. We focus on the differentiation of these proteins into functional and structural related subgroups based on their analysis using supervised classification models.

1.2. Objectives

The objectives of the research reported in this thesis are two-fold. On one hand, we aim to make some contributions in terms of Artificial Intelligence methods, particularly as applied to Data Mining (DM) and ML techniques. These techniques assist the analysis of alignment-free protein sequences and can be useful as well in other application domains. Among the contributions, we may cite new methodologies for detecting mislabeled data or expertise on the use of specific ML methods for the investigation of proteomic data as, for example, in protein classification. In particular, we refer to the construction of multi-class classification models that are able to handle the high dimensional feature sets generated from the amino acid sequences of the proteins. We focus on the objective of evaluating classification models and their optimization through parameter tuning and dimensionality reduction. Another objective is the investigation of feature selection methods for the identification of the most relevant features from the protein classification models, as this information may give insight about the underlying data to the domain expert, i.e. biochemical experts may investigate whether the selected features have a relevant functional or biological significance.

On the other hand, the objective at the biochemical level is to contribute to the research on class C GPCRs, which are of special interest in pharmacology. A first aim is the development of effective and efficient classifiers for discriminating C GPCR subtypes at different levels from their amino acid sequences and/or some data transformations based on physico-chemical features. A second aim is to discover characteristic motifs for each subtype as no specific motifs were known for this class of GPCRs at the beginning of this investigation[2]. The finding of short amino acid patterns, which may be key to distinguishing the several subclasses of class C GPCRs may bring more insight about the specific functionalities of these classes. In order to reach these achievements, it is important to take into account the internal structure of the amino acid sequences, which consists of several transmembrane, intracellular and extracellular regions.

1.3. Methodology

A detailed review of the state of the art regarding the use of ML techniques in the field of proteomics is required as a first step towards these goals, being the main subject the study of methods for classification and feature selection of alignment-free protein sequences. The construction of protein classification models should be approached using different types of supervised classification techniques, but we will focus mainly on the use of Support Vector Machines (SVMs) from labeled data

(e.g. GPCRdb database). This also includes the following subproblems: finding adequate data preprocessing and transformations; selecting relevant features to identify given patterns of substructures appearing in GPCRs. As a result of this, another issue of interest is the extraction of biologically relevant information from the developed classification models, in particular the analysis of subtype specific motifs.

The overall organization of the investigation can be separated in three parts:

- In the first part of the research, the main topic is the construction of supervised classification models for unaligned protein sequences based on physico-chemical transformations and n -gram representations of the sequences. These models are useful to assess the internal data quality of the externally labeled dataset and to handle with the label noise problem.
- In the second part, we work on the analysis of the sequences to discover subtype and region specific motifs. Therefore, we carry out a systematic analysis of the topological sequence segments with supervised classification models in order to evaluate the subtype discrimination capability of each region. In addition, we apply different types of feature selection techniques of the n -gram representation of the amino acid sequence segments to find subtype and region specific motifs.
- Finally, we compare the findings of the motif search, i.e. the region specific motifs with the known 3D crystallographic structures of the class C GPCRs.

1.4. Main thesis contributions

The goal of the thesis was to analyze the class C GPCR dataset systematically with supervised ML approaches. Several different aspects of the proteomic data were analyzed leading to contributions both in the field of Artificial Intelligence and Pharmacoproteomics, listed as follows:

The first contribution (Chapter chapter 4) comprises the construction of robust supervised multi-class classification models for the class C GPCR using several different types of data transformations, which discriminate accurately the subtypes at different levels from their amino acid sequences.

As a second contribution, and based on the detection of a small set of recurrent and persistent misclassifications in the supervised classification models, a novel systematic approach for misclassification analysis is proposed in chapter 5 and exemplified with the class C GPCR dataset. Both the construction of classification models and the analysis of misclassifications hinted about the internal data quality

of the analyzed dataset in reference to subclass separability and correctness of data labels.

The next contribution (chapter 6) takes into account the quantitative and qualitative evolution of the analyzed database over time. The original class C GPCR database from 2011 is compared with two more recent versions of 2016 with special focus on the internal data quality of the respective datasets. The aforementioned novel systematic misclassification analysis approach and supervised classification models are employed to evaluate the internal data quality.

As further contribution (chapter 7), we carried out a complete and detailed analysis of the topological segments of class C GPCRs with regard to their subtype classification capability. The study presents a systematic analysis of the classification performance of the individual sequence segments in which the sequence can be divided in each of its structural domains, as well as the performance of several of their combinations. The result from this study was the identification of the most discriminative segments, which should be the starting point for future work focusing specifically on separate regions. Such future research should involve feature selection starting from these segments as a way to discover specific motifs with subtype discriminative capabilities and potential functional roles.

The results and contributions regarding the discovery of motifs are described in chapter 8, where class C GPCR subtype characteristic amino acid patterns are identified through feature selection. Several feature selection approaches are investigated either on the complete sequence and on the N-terminus segment, which was found to be highly discriminative between subtypes according to the findings of the topological segmentation analysis (See chapter 7). The main contributions of this research in the field of pharmacoproteomics are the identification of subtype discriminating amino acid patterns as well as the identification of subtype characteristic amino acid patterns. In order to filter these different types of patterns, we investigated the use of different multi-class classification methodologies: A subtype-*vs*-subtype classification approach was used to yield subtype discriminative amino acid patterns, while a subtype-*vs*-all the rest of subtypes procedure was used to shift towards the selection of those patterns that distinguish each class C subtype from the rest. With respect to the investigation of feature selection methods, we found a two-step approach combining an univariate t-test filter with a subsequent sequential forward selection especially effective for the discovery of subtype characteristic pattern. The results of this two-step approach were compared with the metrics from a χ^2 filter evaluation, which revealed coincidence in the pattern selection between both approaches.

The last contribution (chapter 9) involves an analysis of the crystal structures of the receptor N-terminus with regard to the subtype specific amino acid patterns

Topic	Chapter	Publication
Construction of supervised classification models using alignment-free transformations	chapter 4	[3]
A systematic approach to GPCR misclassification analysis	chapter 5	[4] [5, 6]
Tracking the evolution of class C GPCR database	chapter 6	[7]
Topological sequence segments discriminate betw. class C GPCR subtypes	chapter 7	[8, 9]
Feature selection for the identification of subtype-discriminating n -grams.	chapter 8	[10, 11] [12]
Feature selection for the identification of subtype characteristic n -grams.	chapter 8	[13]

Table 1.1.: Overview of publications

found during the motif search. Several frequent amino acid patterns identified as potential subtype specific motifs were indeed detected in the analyzed crystal structures. The frequent appearance of these amino acid patterns in the known crystal structures gives foundation to further investigate the significance of these amino acid patterns in the field of computational chemistry as a future line of research.

1.4.1. List of publications

Table 1.1 displays the list of the scientific publications resulting from the investigation of the different topics of the thesis.

1.5. Organization of the thesis

This Section provides the reader with an overview of the structure of this thesis. In the first part, chapter 2 and chapter 3 explain the biological and technical background of the research. The following chapters make up the main research part and comprise the contributions and original research of this thesis, that is, the analyses of the class C GPCR datasets following different supervised ML approaches. In brief:

- Construction of supervised classification models using alignment-free transformations (chapter 4).

- A systematic approach to GPCR misclassification analysis (chapter 5).
- Tracking the evolution of a Class C GPCR database using ML tools for biocuration assistance (chapter 6).
- Topological sequence segments discriminate between class C GPCR subtypes (chapter 7).
- Discovering class C GPCR motifs (chapter 8):
 - Feature selection used for the identification of subtype-discriminating n -grams (8.2).
 - Feature selection used for the identification of subtype characteristic n -grams (8.3).
- Analysis of 3-D crystal structures (chapter 9).

The thesis wraps up with a chapter dedicated to general conclusions and an outline of potential future lines of work.

2. Biological Background

2.1. Introduction

GPCRs are cell membrane proteins with a key role in regulating the function of cells. This is the result of their ability to transmit extracellular signals, which makes them relevant for pharmacology. This has led, over the last decade, to active research in the field of proteomics.

The functionality of a protein mostly depends on its 3-D structure, which determines its ability for certain ligand binding. In 2011, the 3-D structure was only fully determined for approximately a 12% of the human GPCR superfamily [1]. As an alternative, when the information about the 3-D structure is not available, the investigation of the functionality of a protein can resort to the bioinformatics analysis of its primary amino acid (AA) sequence. This information is known and available in several public curated databases.

In this chapter we describe some basic concepts of protein analysis from the field of bioinformatics, such as the different protein structures, the concept of sequence motif and a review of some related public databases. Finally we will see an exposition about the data analyzed in this research, GPCRs in general and specifically its class C, as well as a review of GPCRdb, the here used source of data.

2.2. Proteomics

2.2.1. Protein structures

Proteins are large and complex molecules, which are involved in the regulation of essential cell processes such as the catalysis of biochemical reactions, DNA replication and messaging of internal and external signals.

These are macromolecules built up by at least one large chain of AAs. The constituent AAs of the chain are linked with peptide bonds and their linear sequence is referred to as primary structure. The information about the primary structure

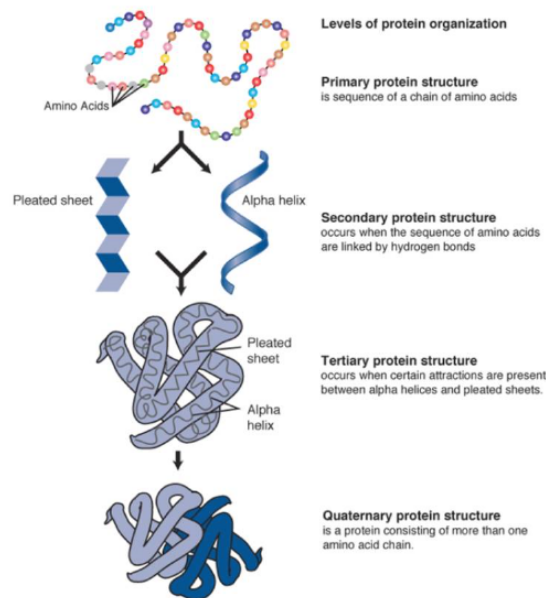


Figure 2.2.1.: Protein structure levels

can straightforwardly be obtained through protein sequencing techniques such as mass spectrometry [14]. The primary sequence therefore constitutes the most common information in the many available protein databases, from which information about its 3-D structure and related interactions with certain ligands can be derived. Many protein sequence databases, such as UniProt, Swiss-Prot, PROSITE or Pfam between others, publish the information about protein sequences and annotate known functionalities.

From the primary structure, the secondary structure can be derived, as hydrogen bonds between amide groups (groups of AA residues) give rise to simple three dimensional patterns which appear repeatedly in the protein backbone, such as alpha helices, beta sheet, loops or coils.

The tertiary and quaternary structures are the 3-D shape of a protein (See Figure 2.2.1) derived respectively from the spatial arrangement of the secondary substructures or complete AA chains, which interact by non-covalent electromagnetic forces between its subunits. The knowledge about the 3-D structure is essential in pharmacoproteomics research as the shape determines the ability of the protein for protein-protein interaction, such as the ability to bind to certain ligands.

Nevertheless, the information about the tertiary or quaternary structure is not always available, as relatively expensive technologies such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy must be applied. Especially

for membrane proteins, such as GPCRs, which are not easily solvable, the information about their full 3-D structure is scarce and alternative methods such as protein structure prediction from its primary structure must be used. In spite of great advances in the field of computational chemistry, protein structure prediction is extraordinarily complicated and still not completely solved, as it requires the calculation of the state of the minimum energy conformation of a protein. At the present, methods such as homology modelling and fold recognition are commonly used to infer information about unknown structures assuming that homolog proteins have similar 3-D structures. The Protein Data Bank (PDB) is a central resource in structural biology that collects experimentally obtained protein structure data throughout the world. The protein structures are acquired from crystallography or similar methods and available in the pdb format, which can be visualized and examined with free available software such as Chimera ¹.

2.2.2. Sequence motifs

Sequence motifs are short linear pattern of AAs that are conserved between a group of proteins and should have a biological significance, i.e. a key function in protein-protein interaction such as a receptor-ligand interaction site, or an enzyme activation site depending on the type of protein. Note that an ideal or *signature* motif, and thus a candidate for potential structural and functional roles, has been described to be one “that matches all the sequences of the target family and no other sequence outside this family” [15]. Motifs are usually expressed as regular expressions using the 20 AA alphabet (See Table 3.3.2). They can be either *contiguous*, if there are no gaps between the AAs that constitute de sequence, or *gap* motifs, if such gaps (filled with any AA of the analyzed alphabet, known as a *wild-card*) are allowed [15].

Known functionalities of these AA patterns are annotated together with the protein structures in the aforementioned publicly available protein databases. It has been suggested that motif over-representation maybe due to evolutionary preservation of sequence segments, indicating their structural and functional roles [15]. For this reason, the search by homology in annotated databases is important to retrieve information about the potential functionality of unknown proteins. Mini Motif Miner (MnM)[16] and Eukaryotic Linear Motif resource (ELM) [17] are motif databases that collect information about known motifs from the literature.

On the other hand, the identification of similar or conserved AA patterns between groups of proteins is important in order to discover family specific motifs. A wide range of bioinformatics tools that use statistical analysis to detect short

¹<http://www.cgl.ucsf.edu/chimera>

overrepresented patterns are available. Multiple EM for Motif Elicitation (MEME) [18], for instance, provides a wide collection of motif discovery tools. PRINTS-S discovers family motifs by sequence alignment of related proteins. Another approach is alignment-based phylogenetic methods, which analyze the evolutionary relation between protein sequences and are able to discover longer conserved AA patterns [19].

2.3. G Protein-Coupled Receptors

GPCRs are proteins located in the eukaryotic cell membrane. This location determines their role as signaling pathway by transmitting extracellular signals to the interior of the cell and thus makes them a prevalent drug target in pharmacological research [20, 21]. Because of their transmembrane location, receptors are able to interact with extracellular signals through activation in the extracellular domain and transmit the signal to the inside of the cell through its effector domain in the intracellular domain experiencing conformational changes [22, 23](See Figure 2.3.1).

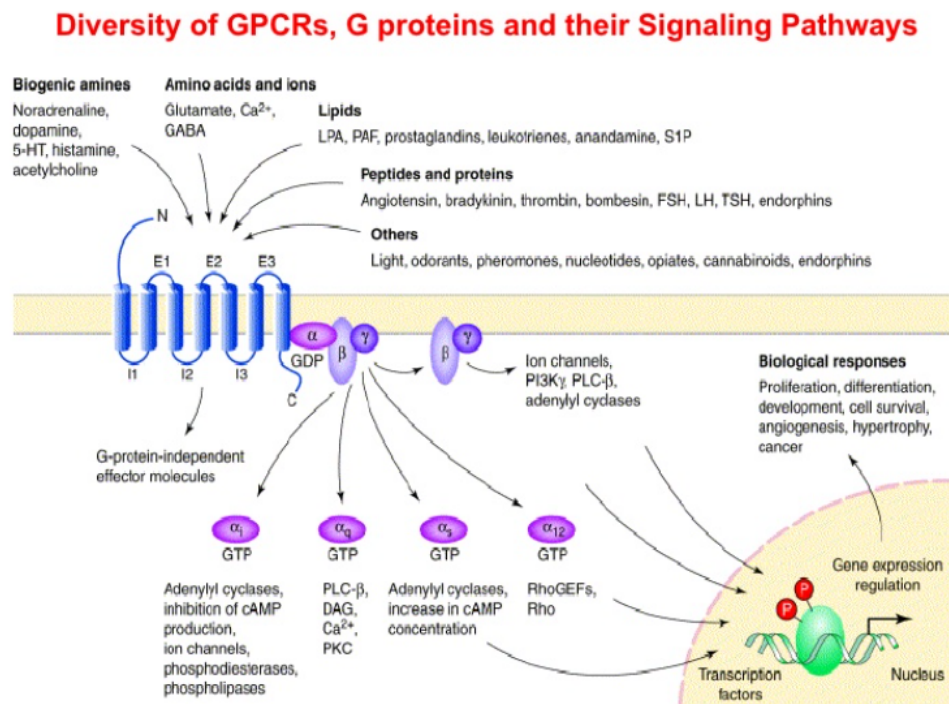


Figure 2.3.1.: GPCR Signalling Pathways taken from [22]

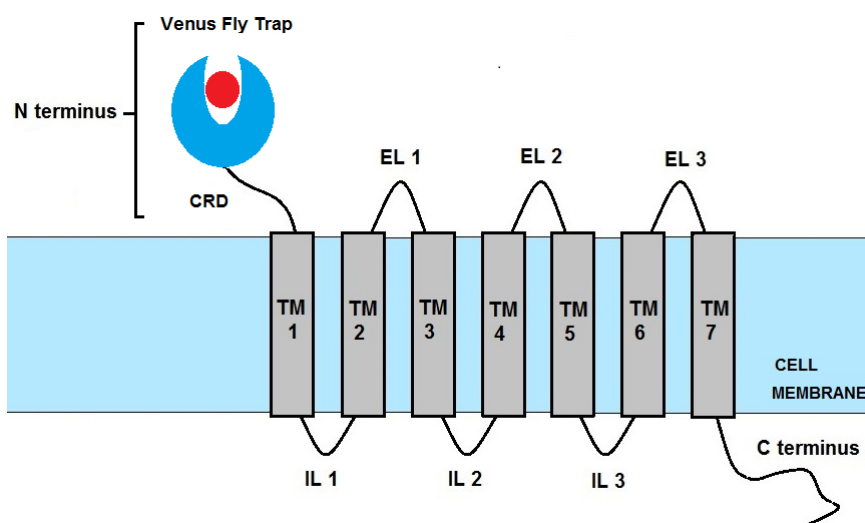


Figure 2.3.2.: Schematic representation of class C GPCRs

The current thesis does not cover the whole of the GPCR super-family. Instead, it specifically targets its class C [24] (defined according to the IUPHAR² convention). Class C GPCR has become an increasingly important target for new therapies, particularly in various central nervous system disorders such as Alzheimer disease, anxiety, drug addiction, epilepsy, pain, Parkinson's disease and schizophrenia [25]. Whereas all GPCRs are characterized by sharing a common seven transmembrane helices (7TM) domain, responsible for G protein activation, most class C GPCRs include, in addition, an extracellular large domain, the Venus Flytrap (VFT) and a cystein rich domain (CRD) connecting both [26, 27]. Figure 2.3.2 shows a schematic representation of the different domains of the GPCR.

GPCR can be activated either at an orthosteric binding site in the extracellular domain or allosterically at the 7TM domain (See Figure 2.3.3). The orthosteric binding site is located in the extracellular domain, specifically at the VFT, which comprises two opposing lobes with a cleft where endogenous ligands bind. Significant efforts are currently devoted by academia and pharmaceutical companies to the design of compounds that, by binding to the 7TM domain, modulate the function of endogenous ligands allosterically, as illustrated by Figure 2.3.3. This multi-domain structural and functional complexity makes class C GPCRs an attractive target for both basic and applied (drug discovery) research. It is worth noting that, although no GPCR allosteric modulators have yet been approved for psychiatric or neurological disorders, a number of GPCR allosteric modulators

²<http://www.iuphar.org>

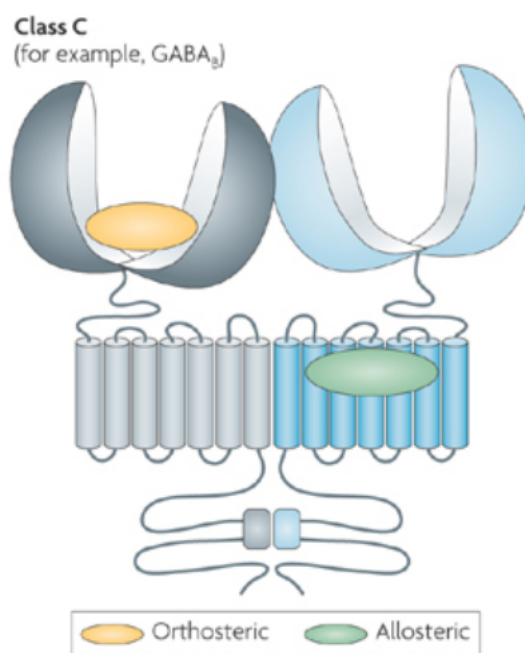


Figure 2.3.3.: Orthosteric and allosteric binding sites.

including, particularly, some from class C, are under clinical development [28]. Allosteric modulators are of especial interest in comparison to orthosteric ligands due to their reduced desensitization, tolerance and side effects, as well as higher selectivity among receptor subtypes and activity depending on the spatial and temporal presence of endogenous agonist [28].

Class C has been further subdivided into up to seven subtypes [29]: Metabotropic glutamate (mG), Calcium sensing (CS), GABA_B (GB), Vomeronasal (VN), Pheromone (Ph), Odorant (Od) and Taste (Ta) receptors. mG receptors are activated by the glutamate amino acid, which is the major excitatory neurotransmitter in the brain; they comprise eight subtypes (mGlu1 to mGlu8) in turn separated into three groups: Group I (mGlu1 and mGlu5), Group II (mGlu2 and mGlu3) and Group III (mGlu4, mGlu6, mGlu7 and mGlu8). Group I mGs signal through G_q whereas Groups II and III signal through G_i/G_o signaling pathways. The mG receptors are involved in major neurological disorders such as Alzheimer and Parkinson diseases, Fragile X syndrome, depression, schizophrenia, anxiety, and pain [30]. It is noteworthy that, although development programs related to the mG drugs *Pomaglumetad* (Lilly), *Mavoglurant* (Novartis) and *Basimglurant* (Roche) for the treatment of schizophrenia, Parkinson disease and Fragile X syndrome have recently been discontinued, some of these drugs are still expected to be beneficial for targeted patient sub-populations with neurological and psychiatric disorders [31].

The CS receptor is activated by the calcium ion and plays a key role in the regulation of extracellular calcium homeostasis. Abnormalities of the extracellular calcium sensing system lead to a disease exhibiting abnormal secretion of parathyroid hormone and hypo- or hypercalcemia. *Cinacalcet* is a marketed positive allosteric modulator of the CS receptor that has proved useful for primary or secondary hyperparathyroidism [25].

The metabotropic GB receptor is activated by GABA, a neurotransmitter which mediates most inhibitory actions in the nervous system. From a structural point of view, the GB receptor distinguishes itself from other class C GPCRs for its lack of CRD. The GB receptor is involved in chronic pain, anxiety, depression and addiction. *Baclofen* is an orthosteric agonist of the GB receptor that is commonly used as a muscle relaxant in multiple sclerosis and as analgesic. Because of their recognized pharmacological advantages, a number of positive allosteric modulators of the GB receptor are currently the goal of programs under development [25].

The investigation of protein functionality and signalling mechanisms is often based on the knowledge of crystal 3-D structures. As previously mentioned, in eukaryotic cell membrane proteins such as GPCRs, this knowledge is partial and fairly recent: The first GPCR crystal 3-D structure was fully-determined in 2000 [32] and over the last decade, the structures of some other GPCRs, most belonging to class A, have been solved [33]. In the case of class C GPCRs the information about tertiary and quaternary structure is very limited, although recent impressive advances in the discovery of GPCR crystal structures [34, 35] have been made. In consequence, the information of the primary AA sequences of class C GPCRs (in this case well known and available from publicly accessible databases) is still required for the investigation of receptor functionality.

2.4. GPCRdb

Pharmacological databases are fundamental for the analysis of the structure and function of biological signal transduction entities, that is, receptors and ion channels [36]. GPCRdb [37, 38] is a web-accessible and publicly-available repository and information system containing data and web tools for GPCR research. Established back in 1993 and currently on its 5th release, it is part of the GPCR Consortium³, an industry-academia partnership and also part of the GLISTEN EU COST Action for the creation of a pan-European multidisciplinary research network.

³URL: <http://gpcrconsortium.org>

GPCRdb characterizes the GPCR superfamily as the union of five major families (namely, A to E) based on functions, ligand types and sequence similarities [39]. Overall, the GPCRdb dataset contains 14,951 proteins from 3,184 species. This resource has been available from 1993 and its management was transferred in 2013 to Prof. David Gloriam’s research group at the University of Copenhagen in Denmark.

The categorization of the receptors available from this database follows the international IUPHAR system recommendations. The whole database originally consisted of seven families: A (Rhodopsin), B1 (Secretin), B2 (Adhesion), C (Glutamate), F (Frizzled), Taste 2 and “other” GPCRs. This classification followed the system suggested in [40].

As previously introduced, the current research focuses only on one of the GPCR families, namely class C, which has become extremely relevant to current pharmacoproteomics research for the selection of some of its members as drug development targets for human central nervous system therapies in areas such as pain, anxiety, or neurodegenerative disorders [24, 41].

Class C in turn comprises several subtypes. At the highest level of grouping, class C discriminates receptors as *ion*, *amino acid*, or *sensory* according to the type of ligand. At the second level of classification seven subtypes are distinguished: metabotropic glutamate receptors (mG, amino acid), GABA_B (GB, amino acid), calcium sensing (CS, ion) and taste 1 receptors (Ta, sensory), covering sweet and umami tastes, as well as also three more sensory-related subtypes of the second level, namely vomeronasal (VN), pheromones (Ph) and odorant (Od) receptors.

At the beginning of this research we used exclusively the dataset from version 11.3.4, as of March 2011 as source about class C GPCR sequences. During the thesis work, GPCRdb underwent major changes in the class C GPCR dataset during 2016, which required an extension of the work to a more recent version of the dataset that does not include some of the aforementioned subtypes.

2.4.1. Class C GPCR datasets

This research focuses mainly on the class C GPCR dataset as of March 2011 (dataset version 11.3.4) published on GPCRdb, but also analyzes two more recent versions published in May and September 2016.

Over the five years elapsed between the earlier and later versions of the database analyzed in this study, GPCRdb has undergone major changes in the total numbers of proteins belonging to class C, but also in the ratio of the different subtypes to the total number of receptors and even in the sequences contained in each of

Subtype	March 2011	May 2016	Sept 2016
mG	351	467	516
CS	48	125	103
GB	208	60	89
VN	344	0	0
Ph	392	0	0
Od	102	0	0
Ta	65	193	228
Total	1510	845	936

Table 2.1.: Number of receptors in each subtype for the class C datasets of the different database versions.

those subtypes. Table 2.1 details the number of sequences for each subtype for the March 2011, May 2016 and September 2016 datasets, respectively (see Figure 2.4.1 for illustration). A mere comparison of the datasets shows a remarkable reduction of the number of sequences, from the 1,510 sequences of the March 2011 dataset, down to the 845 of the May 2016 version and the 936 of the September 2016 one. Moreover, the variety of subtypes included in class C was reduced from the seven of the 2011 dataset to only four in both 2016 datasets. Three receptor subtypes (VN, Od and Ph) were removed in full from class C, but also the number of proteins in the other remaining subtypes changed significantly.

Illustration of evolutionary relationships Figure 2.4.2 displays the evolutionary relationships between the seven sequence subtypes of the 2011 dataset using a phylogenetic tree (PT). A PT is a dendrogram-like graphical representation of the evolutionary relationship between the taxonomic groups that share a set of homologous sequence segments. Specifically, Figure 2.4.2 shows a Treevolution⁴ radial PT plot [42] for the 1,510 GPCR sequences under investigation and their separation into subclasses. This representation provides evidence of the heterogeneity of some of the subfamilies (such as mG, Ph and Od), as they are shown to occupy several different evolutionary branches of the tree. Although less obvious in this particular representation, there is some degree of overlapping between the different subtypes in their tree representation.

⁴<http://vis.usal.es/treevolution/>

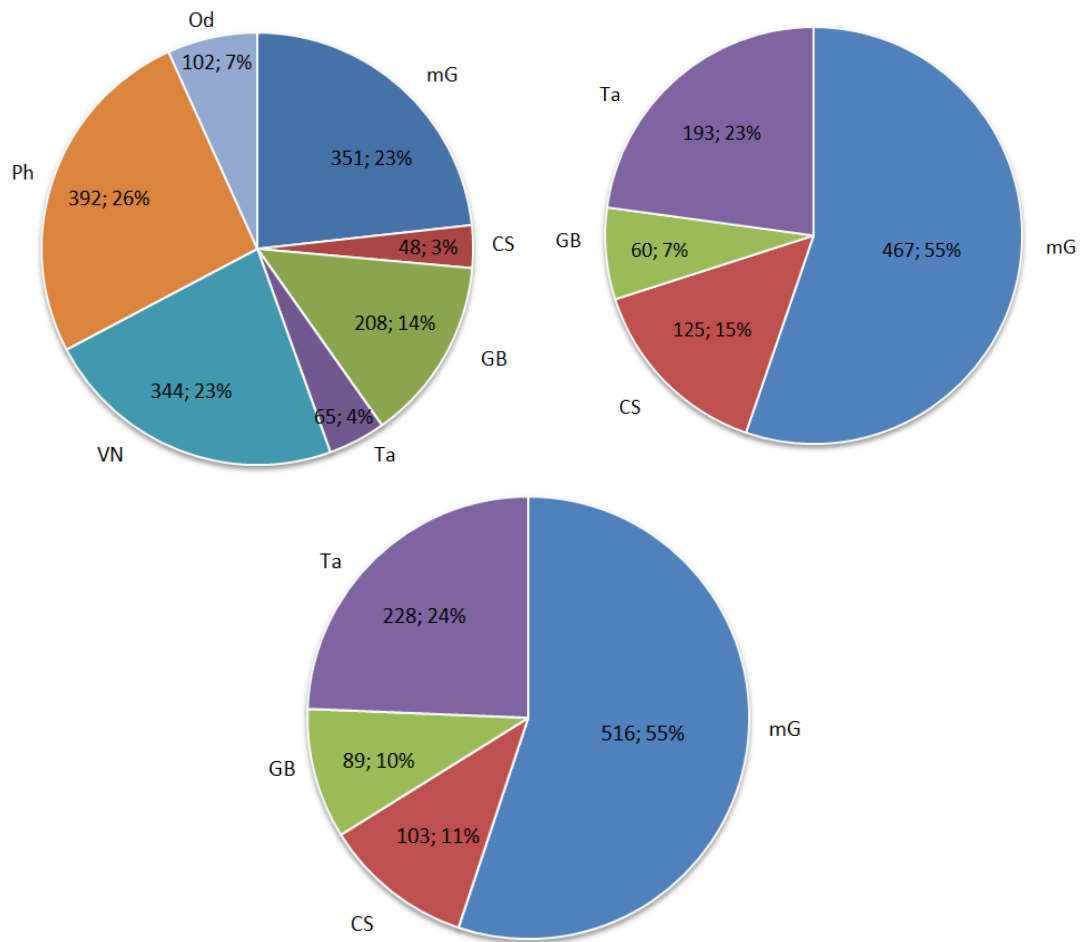


Figure 2.4.1.: Subtype distribution for the different dataset versions (without orphans): upper left - March 2011, upper right - May 2016, middle - September 2016

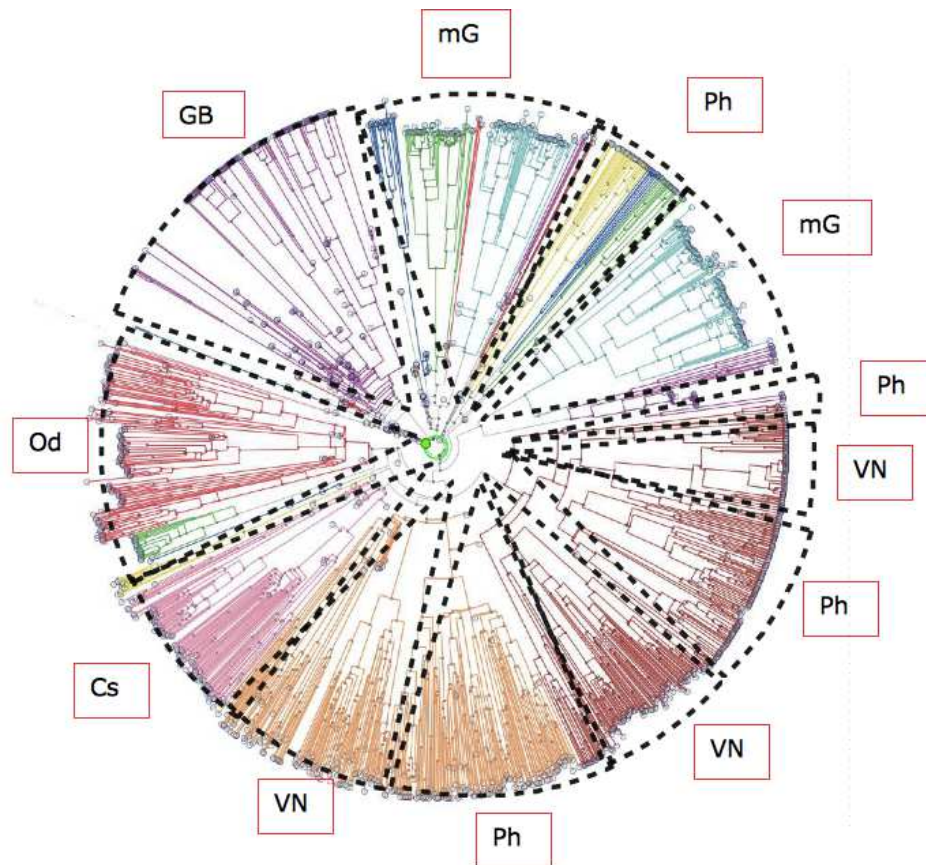


Figure 2.4.2.: Treevolution radial phylogenetic tree for the 1,510 sequences of the 2011 version dataset taken from [6]. Each outer leaf of each branch corresponds to a single sequence, tree colors represent families of descendant nodes. Subfamilies mG, Ph and Od are shown to cover several unrelated evolutionary branches.

3. Technical background

3.1. Introduction

In the field of bioinformatics much research has focused on the analysis of aligned AA sequences. Recently alternative approaches using ML techniques for the analysis of alignment-free sequences have been proposed. In this thesis, we focus on the alignment-free analysis of class C GPCRs sequences using supervised ML methods.

This chapter is thus devoted to the description of the technical background of supervised learning techniques as applied in proteomics research.

We start with the description of supervised learning techniques themselves, including an overview of these methods in proteomics research. This is followed by the description of alignment free data transformations, feature selection and the problem of data label noise.

3.2. Supervised Learning Techniques

In proteomics research, different supervised models for the classification of the alignment-free AA sequences are used for the construction of robust classification models for protein homology detection. SVMs have become commonplace in different problems related to the classification of proteins from their primary sequences. A non-exhaustive list of examples includes SVM-HUSTLE [43], SVM-I-sites [44], SVM-n-peptide [45], and SVM-BALSA [46]. In [47, 48], SVMs were reported to be top-performing techniques for the classification of sequences from similar physicochemical transformations to those used in the current study. In further detail, [49] used the Mean Transformation to classify the five major GPCR classes using Partial Least Square Regression, and [48] used the AA Physicochemical Distance Transformation to classify a benchmark protein database with SVMs. Nevertheless, some studies [50] report better results using more simple models such as Decision Trees (DTs), Naive Bayes (NB), or Random Forests (RF), to name a few. For this reason, a set of different supervised classifiers are compared to the reference SVMs in the current thesis.

3.2.1. Classification Models

SVMs [51] are complex classifiers with an ability to find a linear separation of instances in a higher dimensional space. DTs [52] predict class membership by examining the discriminative power of the attributes. NB is a simple probabilistic classifier that applies Bayes' theorem under the assumption of attribute independence, creating a probabilistic model for class prediction. RF [53] is an ensemble based learning method [54] in which each of the elements of the ensemble is a decision tree [52] and the classification decision is the result of an internal voting system. All these models are described in more detail next.

Naive Bayes

NB [55] is a relatively simple model that provides a baseline for comparison. It is a probabilistic classifier that applies Bayes' theorem with an assumption of independence of variables. Under this assumption the probability of a class given the input data is expressed as $P(C_i|X) = P(C_i) \prod_{n=1}^N P(X_n|C_i)$. This probability could be used for class prediction using Maximum a Posteriori (MAP) estimation in the form $y = \arg \max_i P(C_i) \prod_{n=1}^N P(X_n|C_i)$. The classifiers differ depending on the assumption about the probability distribution for $P(X_n|C_i)$. For continuous variables, the typical assumption is a Gaussian of the form:

$$P(X_n|C_i) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp \left\{ -\frac{(X_n - \mu_n)^2}{2\sigma_n^2} \right\} \quad (3.2.1)$$

The parameters μ_n and σ_n are estimated using a Maximum Likelihood approach.

Random Forest

RF [53] is an *ensemble learning* method [54] using DT-based classifiers [52]. The DT classifiers are trained to split an input space into homogeneous regions with associated class labels. The splits are typically axis-aligned and are selected to maximize the *information gain*.

The main improvements of RF over a single DT are the usage of the bagging technique and the random subsampling method, which both allow avoiding overfitting. Bagging (or bootstrapped aggregating) is a technique of model averaging that uses models trained on subsamples of the original training set. The subsampling is performed independently with replacement.

Support Vector Machines

These methods have their foundations on statistical learning theory and were first introduced in [51].

They map the D -dimensional vectors $\mathbf{x}_i, i = 1 \dots N$, where $\mathbf{x}_i \in \mathbb{R}^D$ and N is the number of instances, into possibly higher-dimensional feature spaces by means of a function ϕ .

The goal is finding a linearly-separating hyperplane that discriminates the feature vectors according to class label with a maximal margin, while minimizing the classification error ξ .

The most simple version is the linear SVM, where a linear hyperplane that separates the examples from two classes is assumed to exist. Such hyperplane is defined by a set of points \mathbf{x} that satisfy $\mathbf{w} \cdot \mathbf{x} - b = 0$, where \mathbf{w} is a normal vector to the hyperplane and $\frac{b}{\|\mathbf{w}\|}$ is the perpendicular distance from the hyperplane to the origin. In consequence, the SVM algorithm, when searching for the hyperplane with largest margin, assumes that $y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0, \forall i$, where y_i are the class labels. The objective of the SVM algorithm is finding the separating hyperplane that satisfies this expression while minimizing $\|\mathbf{w}\|^2$. This problem can be translated to a Lagrange formulation in which the following objective function L_p (primal Lagrangian) must be minimized with respect to \mathbf{w} , b and subject to the restriction that all $\alpha_i \geq 0$:

$$L_P \equiv \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^l \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^l \alpha_i \quad (3.2.2)$$

This is equivalent to the maximization of the dual Lagrangian form L_D :

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \quad (3.2.3)$$

subject to the restriction that all $\alpha_i \geq 0$ and $\sum_i \alpha_i y_i = 0$.

A modification of the algorithm was introduced in [56], allowing a so-called “soft-margin” to account for mislabeled data when a linear separating hyperplane could not be found. A classification error ξ is admitted and a parameter C controlling the trade-off between those errors and margin maximization is defined (Note that, for $C \rightarrow \infty$, the model becomes equivalent to a hard-margin SVM).

The SVM can be extended to nonlinear classification [57] by applying the so-called kernel trick [58]. The use of nonlinear kernel functions allows SVMs to

separate input data in higher-dimensional feature spaces in a way they would not be separable with linear classifiers in the original input space. The use of kernel functions allows to solve the problem without explicitly calculating the mapping ϕ (that is, without calculating data coordinates in the implicit feature space). This is possible due to the following property: $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$, which means that any dot product in the optimization procedure can be replaced by a nonlinear kernel function k . In this thesis, we use the radial basis function (RBF) kernel, specified as $k(\mathbf{x}_i, \mathbf{x}_j) = e^{(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)}$, which is a popular nonlinear choice for SVMs and has been used in the experiments reported in the thesis. With it, the model requires adjusting two parameters through grid search: the error penalty parameter C and the γ parameter of the RBF function, which regulates the “space of influence” of the model support vectors and, therefore, controls overfitting.

SVM Model Selection The construction of SVM classification models using the RBF kernel requires adjusting two parameters through grid search: the error penalty C and the γ parameter of the kernel.

The SVM classification models involve the following processing steps in our experiments:

1. Preprocessing of the dataset: Standardization of the data so that the mean is 0 and standard deviation is 1.
2. Splitting of the dataset into 5 stratified folds and applying 5-fold cross validation (5-CV) for the following steps:
 - a) Use the current training set for a parameter grid-search varying the parameters C and γ in a given range.
 - i. For each combination of C and γ , determine the average classification accuracy using an inner 5-CV and update the parameters C and γ providing the best result.
 - ii. Train an SVM model using the selected parameters C and γ and the current training set.
 - b) Classify the current test set with the SVM model obtained in the previous step recording the classification metrics.
3. Calculate the mean value of the classification metrics recorded during step 2.b over the five outer iterations.

In our experiments, we measure the accuracy and the Matthews Correlation Coefficient (MCC) at the global level and the precision, recall and MCC at class level. The reported measures are the mean values of the respective metric over the five

iterations of the (outer) 5-CV. Further details of the measures are provided below. At each iteration, the aforementioned metrics are recorded for the SVM trained with the best parameters C and γ found in the corresponding grid search. The grid search is conducted in a range of 2^{-15} - 2^5 (step $\times 2$) for the γ parameter and in the range of 1 to 16 (step $+1$) for the C parameter.

SVM multiclass approach The discrimination into several classes, requires extending the original binary (two-class) classification approach of SVMs to a multi-class one. In our experiments, we mainly used the “one-against-one” approach to build the global classification model, which is implemented as part of the LIBSVM¹ library [59]. This approach performs class prediction according to the results of a voting scheme applied to the binary classifiers, i.e., according to the number of times a class is predicted in each binary classifier. Therefore, this multi-class classifier internally uses $K(K - 1)/2$ binary classifiers for distinguishing K classes. A total of 21 binary classifiers are thus built for the experiments with 7 class C GPCR subtypes. Depending on the type of experiment, we will also make use of the “one-against-all” approach, which performs a binary classification for each subclass.

3.2.2. Metrics

Two different figures of merit were used to evaluate the test performance of the multi-class trained classifiers, namely the accuracy (Accu), which is the proportion of correctly classified instances, and the MCC, which involves all the elements of the confusion matrix and it is therefore considered a more complete figure of merit and is most robust for unbalanced datasets [60, 61]. Being the correlation coefficient between the observed and the predicted classification, its value ranges from -1 to 1, where 1 corresponds to a perfect classification, 0 to a random classification and -1 to complete misclassification.

In our experiments, we measure the precision, recall and MCC at class or subtype level (i.e. at the level of the binary classifier) and measure the accuracy and MCC at the global level (i.e., at the level of the multi-class classifier). All these figures of merit, described in Tables 3.1 and 3.2, are based on the concept of true and false predictions in binary classification with “positive” and “negative” classes. True positives (tp) and true negatives (tn) are correctly classified cases of, in turn, the positive and negative classes. Accordingly, false positives (fp) and false negatives (fn) are misclassified cases of, in turn, the negative and positive classes.

¹<https://www.csie.ntu.edu.tw/~cjlin/libsvm/>

Table 3.1.: Performance measures for binary classifiers.

Measure	Formula	Meaning
Accuracy	$\frac{tp+tn}{tp+fn+fp+tn}$	Measure of correctness
Precision	$\frac{tp}{tp+fp}$	Measure of quality
Recall	$\frac{tp}{tp+fn}$	Measure of completeness
MCC	$\frac{tp*tn-fp*fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}}$	Correlation coefficient

Table 3.2.: Performance measures for multi-class classifiers. tp_i , tn_i , fp_i and fn_i are, in turn, tp , tn , fp and fn for class i [62]. The multi-class MCC is calculated taking into account all the entries of the confusion matrix $C_{K \times K}$ involving all K classes[63]. The ij -th entry (c_{ij}) is the number of examples of the true class i that have been assigned to the class j by the classifier.

Measure	Formula
Accuracy	$\frac{\sum_{i=1}^K \frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}}{K}$
MCC	$\frac{\sum_{k,l,m=1}^K C_{kk}C_{ml}-C_{lk}C_{km}}{\sqrt{\sum_{k=1}^K [(\sum_{l=1}^K C_{lk})(\sum_{f,g=1,f \neq k}^K C_{gf})]} \sqrt{\sum_{k=1}^K [(\sum_{l=1}^K C_{kl})(\sum_{f,g=1,f \neq k}^K C_{fg})]}}$

By using a 5-fold cross-validation (CV) procedure to evaluate the multi-class trained classifier, the reported measures are the mean values of the respective metrics over the five iterations of the 5-CV.

3.2.3. Cross Validation

Although the supervised learning techniques differ in the employed learning strategy, the training and validation approach is common. In our experiments, we apply CV [64], which uses two separated set of examples, one to train the classifier and another to validate its quality of prediction.

More generally, in CV the entire dataset is split into a given number n of stratified folds, where the percentage of class membership of each class are preserved in the folds. The class prediction function is inferred from the training set (formed by the instances from $n-1$ of the n folds) and its quality of prediction is validated classifying the instances from the test set, which is built by different examples as from the training set. A common setting is the 5-fold cross validation (5-CV), where 4 folds are used for training and the 5th fold for the validation.

CV uses an iterative approach to infer and validate the classification function, where the described cross validation approach is repeated n times and the classification metrics are reported as the mean value of all iterations.

3.3. Alignment-free data representation

As the AA sequences of the proteins have a variable length, one might apply sequence kernels [65, 66, 67] or transform the sequence data to fixed-size vectors in order to use them with any supervised classifier, including non-kernel methods such as DTs and NB. Here, we follow the latter approach and transform the AA sequences to fixed-size vectors. In the following, we describe the different transformation methods applied to the analyzed class C GPCR datasets. We distinguish between methods based on the physicochemical properties of the AAs and those based on the n -gram representation built from the AA alphabet.

3.3.1. Transformations based on the physicochemical properties

In this thesis, we decided to use several distinct transformations based on the physicochemical properties of the AAs and the sequencing information such as

Auto-Cross Covariance (ACC) [68], the Mean Composition [49] and Physicochemical Distance-Based Transformation (PDBT [48]). Beyond computational convenience, the use of transformations based on the physico-chemical properties of the AAs is justified by the fact that, as stated in [48], “because protein structure and function are more conserved during evolutionary process, the similarity between two distantly related proteins may lie in the physicochemical properties of the AAs rather than the sequence identities”. In the following, we describe each of the transformations in some detail:

- **Auto Cross Covariance Transformation:** The ACC [69, 68] is a sophisticated transformation, capturing the correlation of the physico-chemical descriptors along the sequence. First, the physico-chemical properties are represented by means of the five z -scores of each AA, as described in [70]. Then, the Auto Covariance (AC) and Cross Covariance (CC) variables are computed on this first transformation. These variables measure, in turn, the correlation of the same descriptor (AC) and the correlation of two different descriptors (CC) between two residues separated by a lag along the sequence. From these, the ACC fixed length vectors can be obtained by concatenating the AC and CC terms for each lag value up to a maximum lag, l . This transformation generates a $N \times (z^2 \cdot l)$ matrix, where $z = 5$ is the number of descriptors. In this work we use the ACC transformation for a maximal lag value of $l = 13$, which was found in [71] to provide the best accuracy for the analyzed data set.
- **Physico-chemical Distance-Based Transformation:** The PDBT transformation [48] is a complex transformation that uses a large set of physico-chemical properties: 531 values representing physicochemical and biochemical properties of AAs are taken into account. Furthermore, sequence-order information is incorporated in the representation in the form of the correlation of each property between two AAs separated by a maximal lag l . In the current study, we use the PDBT transformation for a maximal lag of 8, which yields a $N \times 4248$ matrix that was previously analyzed in [72].
- **Mean Composition Transformation:** This transformation applied in [49] first translates the AA sequence into physico-chemical descriptions, i.e. each AA is described by five z -scores [70]. In order to obtain a fixed-length representation of the sequence the average value of each z -score is calculated. This transformation generates a $N \times 5$ matrix.

AA name	Symbol	AA name	Symbol
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	Tryptohan	W
Histidine	H	Tyrosine	Y
Isoleucine	I	Valine	V

Table 3.3.: List of elements of the 20 Amino acid alphabet.

3.3.2. N-gram representations

The use of the n -gram representation is common in protein characterization and has been investigated in, for instance, [73, 74, 75]. This type of transformation has its foundations in the field of symbolic language analysis. They treat protein sequences as text from the 20 AA alphabet [74, 76] (See Table 3.3.2). Here, the appearance of short sequence fragments known as n -grams, are understood as “words”. In [50], a successful application of class A GPCR classification using text classification methods was reported using a discretization of n -gram features. In this research, we followed a similar strategy and calculated the relative frequency of occurrence of n -grams of sizes one and two that we call, in turn, AA and Digram transformations.

- **N -gram representations:** These transformations partially disregard sequential information to reflect only the relative frequency of appearance of AA subsequences. In the case of AA, the frequencies of appearance of the 20 AAs (1-gram) are calculated for each sequence (i.e., a $N \times 20$ matrix is obtained, where N is the number of items in the dataset). In the case of the Digram (2-gram) method, we calculate the frequency of each of the 400 possible AA pair combinations from the AA alphabet (i.e., a $N \times 400$ matrix is obtained). N -grams can be understood as receptor sequence deterministic motifs. They can be either *contiguous*, if there are no gaps between the AAs that constitute de n -gram, or *gap* motifs, if such gaps (filled with any AA of the analyzed alphabet, known as a *wild-card*) are allowed [15].

We also use more complex natural language processing (NLP)-inspired transformations such as the Prot2Vec distributed transformation:

GROUPING	1	2	3	4	5	6	7	8	9	0	X
SEZ	IVLM	RKH	DE	QN	ST	A	GT	W	C	YF	P
DAV	SG	DVIA	RQN	KP	WHY	C	LE	MF	T		

Table 3.4.: Amino acid grouping schemes.

- **Prot2Vec** distributed transformations: This transformation has its foundations in NLP. To apply this method to protein sequence classification, the AAs are considered as letters and the whole sequences as sentences, with n -grams acting as words. In NLP, this representation is understood as “distributed” because one “concept” in the domain is represented in several dimensions and one dimension gathers information about several “concepts”. In NLP, these distributed word representations are learnt using an Artificial Neural Network model and have been refined in the form of Continuous Bag-of-Words (CBOW) and Continuous Skip-Gram (CSG) models [77]. This idea was extended to protein sequences in [78], where it was shown to capture meaningful physical and chemical properties of the proteins. In this study based on the research of [79], 3-gram representations were first created from two different databases: Swiss-Prot and GPCRdb. The GPCRdb representation was created using the complete database (not only class C). To train the model, each sequence was split into 3 sequences of 3-grams with offsets from 0 to 2. Each of these 3 sequences were used in training set. A skip-gram version of window size 25 was used to train both models. For the final working representation of a sequence, the vectors corresponding to its 3-grams were summed up. We refer as Prot2Vec1 or Prot2Vec2 transformation to that created from the Swiss-Prot or GPCRdb database respectively .

Amino acid alphabets

Commonly, n -gram representations use the 20 AA alphabet (See Table 3.4), but also different groupings of AAs may be used. According to [80], many AAs have similar physicochemical properties, which makes them equivalent at a functional level. An appropriate grouping of AAs reduces the size of the alphabet and may decrease noise. In this work, besides the basic 20 AA alphabet, we used two alternative AA groupings (See Table 3.4): the Sezerman (SEZ) alphabet, which includes 11 groups, and the Davies Random (DAV), including 9 groups. They have both been evaluated [81] in the classification of GPCRs into their 5 major classes.

N-gram length	AA	SEZ	DAV
1	20	11	9
2	400	121	81
3	8,000	1,331	729
4	160,000	14,641	6,561

Table 3.5.: Size of n -gram feature spaces.

3.4. Feature Selection

In many bioinformatic applications, feature selection (FS) approaches are commonly used to discard irrelevant and redundant features [82]. This also happens for the n -gram frequency representation of protein data [50], where longer n -gram patterns may be derived from the sequence leading to feature spaces, which grow exponentially with the lengths of the n -gram pattern and number of elements in the alphabet. Table 3.5 shows the size of the n -gram feature space for the AA, SEZ and DAV alphabets used in this study (See Table 3.4).

In the context of supervised classification, the n -gram frequency representation of higher lengths becomes intractable, both because the algorithm was not designed to manage vast feature spaces or because of overfitting of the classifier model and the consequent loss of generalization. FS increases the quality and interpretability of the classification models as they select a reduced subset of relevant features, which may give insight about the underlying data to the domain expert. A difference to other dimensionality reduction methods based on feature extraction, such as, for instance, principal component analysis, FS techniques select a subset of unaltered observed features.

FS comprises filtering techniques, wrapper methods and embedded methods:

1. Filtering techniques: The feature's relevance is evaluated by a univariate metric (χ - *square*, t-test, Euclidean distance, Gain ratio, etc.), or a multivariate metric (Markov Blanket filter [83], Correlation-based feature selection, etc.). The highest scoring features are selected as candidate set for any classifier. This approach is fast, scalable and independent from the classifier, but with the drawback that the univariate metrics may ignore feature dependencies. [84] presents a comparative analysis of several filter methods combined with a diverse range of supervised classifiers.
2. Wrapper methods: The feature subset search is combined with the classification model search. Deterministic approaches comprise sequential forward selection or sequential backward elimination [85]. Greedy algorithms such as Simulated Annealing, Randomized Hill Climbing [86] or Genetic algorithms

[87] are more randomized and less prone to select local optima, but computationally more expensive than the deterministic approaches. Wrapper methods perform feature selection dependent on the type of classifier.

3. Embedded methods: The classifier algorithm performs feature subset selection as part of the model construction, measuring which features contribute best to the classification performance. Internally, these methods use penalizations to discard irrelevant features. DTs [88] and weighted vectors from SVMs or LASSO regression [89] are some examples. [90] describes a recent application of FS with embedded methods for a mass spectrometry data set.

In n -gram representations in proteomics, we deal with a vast feature space. In those cases, a two-step FS approach or a combination of several methods may be applied to reduce the feature space appropriately [91]. In this thesis, we propose a two-step FS approach consisting of a t-test or χ^2 filter as first step, followed by a sequential FS algorithm to search for n -grams which discriminate best between the class C GPCR subtypes. Later on, these three approaches will be explained in more detail.

3.4.1. t-test Filtering

Two-sample t-tests are a coarse evaluation of the discriminating power of individual features (n -gram frequencies). This univariate statistical test analyzes whether there are foundations to consider two independent samples as coming from different populations (normal distributions) with unequal means by analyzing the values of the given feature. An appropriate significance level must be established in order to decide whether the hypothesis is accepted or not. For instance, with a significance level of 0.01, the hypothesis is accepted when the p -value is below 0.01. If the t-test suggested that this hypothesis was true (i.e. the null hypothesis was rejected), the feature would be considered to significantly distinguish between the two different subtypes of class C GPCRs. The p -value of a test measures the *false positive rate* that is assumed to happen considering the test as significant. A false positive is the case when the null hypothesis is rejected when it is really true.

In the case of vast feature spaces when many single t-tests are applied, a complementary metric, the q -value, was proposed for evaluating the significance of the many simultaneous performed t-tests [92]. The q -value of a test measures the *false discovery rate* that happens when considering the test as significant. The false discovery rate is the expected proportion of false positives among the tests found to be significant. This method was applied for the evaluation of the significance of features in genomics data tackling a large number of features [93].

3.4.2. χ^2 Filtering

The χ^2 filtering approach is usually employed in text document classification with n -gram representation. This metric was already successful in a GPCR family classification [50] for n -gram FS. It evaluates the power of the n -gram to classify an instance into two binary classes c by measuring the lack of independence between the n -gram x and the category c . The measure takes into account the number of expected instances having the feature in class c $e(c,x)$ and the number of instances of class c actually having the feature $o(c,x)$. The expected number of instances $e(c,x)$ is calculated assuming independence of the term x from the class c according to the formula $e(c,x) = n_c \cdot \frac{N}{t_x}$, where N is the total number of instances, n_c is the number of instances in class c and t_x is the number of instances having the feature x . The χ^2 value for feature x is calculated as follows:

$$\chi^2(x) = \sum_{c_i \in C} \frac{[e(c,x) - o(c,x)]^2}{e(c,x)} \quad (3.4.1)$$

3.4.3. Sequential Forward Selection

The second dimensionality reduction step starts from the selection performed through a filtering method (t-test or χ^2 approach) and involves a sequential forward selection algorithm [85]. This algorithm is used to find the reduced set of features that best discriminated the data subtypes. This kind of algorithm is a so-called wrapper method, where the classification model search is performed within the subset feature search [82].

The algorithm starts from an empty candidate feature set and adds, in each iteration, the feature which most improves the accuracy (i.e., that which minimizes the misclassification rate) in a 5-fold cross-validation (5-CV).

3.5. Analysis of Label Noise

ML is a data-driven process and, as such, the quality of the available data is paramount. Label noise (LN) may become a data quality problem in supervised ML and Computational Intelligence and is commonplace in real-world applications [94].

There are few domains of knowledge in which the effects of label noise are so pervasive and eloquent as in biomedicine and bioinformatics [95]. It can take many

forms, including expert subjectivity in the labelling process, bounds on the available information and communication noise [96]. In bioinformatics, protein subtype characterization is riddled with this problem. In the specific area of GPCRs, this problem is magnified by the fact that subtyping can be performed at up to seven levels of detail [97]. GPCR subtype discrimination may use aligned (through Multiple Sequence Alignment, or MSA [98]) or unaligned [99] versions of the sequences, but as a computer-based automated classification procedure it is susceptible to data labeling issues.

3.5.1. Label noise detection

In ML, where regular data distributions are searched which best explain the data, instances which are abnormal to such distributions are known as outliers, i.e. special cases distant from any regular data distribution. Sometimes, what appears as an outlier may just correspond to errors in the data labelling. Often, expert knowledge is required to decide on this issue. In supervised learning, label noise is an important problem. Several filtering methods have been proposed to filter out noisy data instances. A classical label noise detection method is the Classification Filter [100], where several different classifiers are used to filter out noisy instances, i.e. an instance is considered as noise when it is classified incorrectly by the majority of the used classifiers. Ensemble Filters [101] use a similar approach, although a different importance is given to the different classifiers. Their classification errors are combined to detect mislabeled instances using either a consensus vote filter (all classifiers detect a classification error), or a majority vote filter (the majority of classifiers detect an error). A recent contribution in this field is the Noise Rank framework [102], which uses an ensemble filter to filter out noisy instances and classifies them either as outliers or possible errors. Another way to detect label noise are the Saturation Filter methods [103]. These methods search for saturated training sets, which correspond to data sets which best explain the underlying concepts of the data by eliminating those data instances from the training set which increase the complexity of the explanation of the data (also referred to as the Complexity of the Least complex correct Hypothesis measure of the dataset).

All these filters are commonly used in label noise detection in different application domains, such as the medical domain [104], gene expression data [105] or for software development quality data [106].

4. Construction of supervised classification models using alignment-free transformations

Class C GPCR subtype discrimination is addressed here as a supervised classification problem in which class labels are the assignments of each of the sequences to one of the seven existing subtypes according to the database available information. In this chapter, we report the experiments carried out using alignment free data transformations on the complete sequences. We analyze both the dataset version from 2011 and the final version from 2016. The earlier dataset version from May 2016 is not analyzed because of its similarity to the final version of 2016.

4.1. Experiments with the 2011 dataset

4.1.1. Experimental settings

Methods In this section, we analyze the use of different alignment-free transformations of the complete sequence on the class C GPCR dataset published on March 2011 on GPCRdb, comprising 1,510 sequences. We analyze both transformations based on the physicochemical properties of the sequences and n -gram transformations. The first experiments concern the transformations based on the physicochemical properties: we compare the Mean, ACC and PDBT transformations as explained in section 3.3.1. The second set of experiments deals with n -gram transformations from different AA alphabets (AA, SEZ and DAV) as explained in section 3.3.2. We consider n -grams of lengths one and two (1-gram and 2-gram respectively), the combination of both (1-2-grams) and the Prot2Vec distributed transformation.

Both experiments follow a common approach:

1. The first phase of the experiments involves the use of several multiclass classifier models for the classification (SVM, RF and NB). The classification models are built using 5-CV with stratified folds. Regarding the use of

SVMs, the discrimination of the seven subtypes of class C GPCRs requires extending the original binary (two-class) classification approach of SVMs to a multi-class one. To that end, we chose the “one-against-one” approach to build the global classification model, implemented as part of the LIBSVM library. Therefore, this multi-class classifier internally uses $K(K - 1)/2$ binary classifiers for distinguishing K classes. A total of 21 binary classifiers are thus built for the seven class C GPCR subtypes in our study. The classification results of the different classifiers are used to choose which classifier is most adequate for the rest of analyses.

2. In the second phase, the classification model which most accurately classified the dataset is analyzed in detail, namely the SVM model. The global classification results and the per-subtype classification results are reported.
3. In the third phase, we analyze the subtype classification results in more detail focusing on the type of classification errors for each subtype.

The classification performance of the results is measured by means of the accuracy (Accu) and Matthews correlation coefficient (MCC) for multiclass classification. The subtype results are evaluated by means of the MCC, precision and recall for binary classification. All experiments are conducted using 5-CV with stratified folds. SVM classification uses the LIBSVM library and an RBF kernel, whose parameters C and γ are tuned through a grid search.

4.1.2. Experiments with transformations based on the physicochemical properties

In this section, we report the experiments carried out on the class C GPCR dataset using supervised classification methods and transformations based on the physicochemical properties of the sequences.

Results and Discussion

Table 4.1 shows the classification performance measured by Accuracy and MCC for the three transformed datasets (MEAN, ACC and PDBT) obtained by a SVM, RF and NB classifier. Regarding classifier selection, SVM clearly outperforms DTs and NB for all three datasets and therefore should be used in the rest of analyses.

Table 4.2 shows in detail, for the three transformed datasets, the SVM model parameters, which were found in the grid search conducted to find the optimal parameters C and γ of the RBF-SVM following the approach explained in Section 3.2.1. For each dataset, the combination of parameters found to have the best

Table 4.3.: Subtype classification result obtained by SVM from the ACC transformation of the 2011 version dataset.

Subtype	MCC	Precision	Recall	Type I error	Type II error
mG	0.956	0.945	0.988	low	-
CS	0.933	1.000	0.877	-	high
GB	0.986	0.990	0.985	-	-
VN	0.893	0.912	0.924	medium	medium
Ph	0.864	0.896	0.903	high	medium
Od	0.799	0.889	0.744	high	high
Ta	0.991	1.000	0.984	-	-

performance, and the corresponding mean accuracy and MCC values on the test sets, are reported.

Table 4.1.: Accuracy (Accu) and MCC according to dataset and classifier. N stands for the size of the feature set.

		SVM		RF		NB	
DATA	N	Accu	MCC	Accu	MCC	Accu	MCC
MEAN	5	0.68	0.59	0.67	0.58	0.58	0.46
ACC	325	0.93	0.91	0.79	0.74	0.84	0.80
PDBT	4248	0.92	0.90	0.82	0.77	0.71	0.64

Table 4.2.: Model selection results (MEAN, ACC and PDBT).

DATA	PARAMETERS	Accu	MCC
MEAN	$C=2, \gamma=1$	0.68	0.59
ACC	$C=[2,8], \gamma=2^{-9}$	0.93	0.91
PDBT	$C=4, \gamma=2^{-12}$	0.92	0.90

The best classification results are found for the ACC transformed dataset using SVM classifiers (see Table 4.1 for a summary), achieving an accuracy of 0.93 and an MCC value of 0.91. This result obtained for the ACC transformed datasets is consistent with that obtained with semi-supervised techniques in [72], where the ACC dataset outperformed the other transformed datasets.

Table 4.3 shows the classification results for the ACC-transformed dataset and the SVM classifier with greater detail at the per-class level (these results correspond to a model with parameters $C=2$, $\gamma=2^{-9}$). The MCC value shows that classes mG, CS, GB and Ta are very accurately discriminated from the other classes, having an MCC between 0.93 and 0.99. The prediction power of the classifier for classes VN, Ph and Od is clearly lower, with MCC values that range from 0.79 to 0.89.

As for the quality of the classifier, measured by the precision, it can be seen that it provides the most exact results for classes CS, GB and Ta, as its precision gets very close to its maximum possible value. This metric shows that for classes mG, VN, Ph and Od some type I classification errors (false positives) happen. Regarding the completeness of the classifier, measured by the recall, we see that it is most complete for classes mG, GB and Ta, which means that nearly all real positives are correctly predicted. Classes CS, VN and Ph have a lower recall, meaning that some type II errors (false negatives) happen for these classes. Class Od has a significantly lower recall than the other classes, what means that this class is most difficult to recognize.

Table 4.3 also shows an estimation of the quantity of type I and type II errors for each class. An analysis of these errors, by means of the confusion matrix, shows that the type II errors occur recurrently with a specific pattern for each class. For example, Ph are most frequently misclassified as Vn and less frequently as mG or Od. The existence of those patterns in the type II errors encourage an analysis of the class C dataset at the biochemical level in future work.

4.1.3. Experiments with n -gram transformations

In this section, we report the experiments carried out on the class C GPCR dataset using supervised classification methods and n -gram representations, including different alphabets (AA, SEZ and DAV) and more sophisticated transformations such as the Prot2Vec distributed transformations. For a detailed explanation we refer the reader to Section 3.3.2.

Results and Discussion

First, we built classification models with n -grams for each of the three alphabets (AA, SEZ, DAV) (See Table 4.4) and two variants of the Prot2Vec distributed transformations (See Table 4.5).

Table 4.4.: N -gram classification results for the different alphabets, where N is the size of a feature set.

			SVM		RF		NB	
ALPH	DATA	N	Accu	MCC	Accu	MCC	Accu	MCC
AA	1-gram	20	0.880	0.840	0.807	0.751	0.720	0.650
	2-gram	400	0.932	0.914	0.815	0.766	0.834	0.792
	1-2-gram	420	0.934	0.917	0.805	0.753	0.828	0.785
SEZ	1-gram	11	0.815	0.763	0.754	0.686	0.603	0.500
	2-gram	121	0.923	0.903	0.833	0.788	0.805	0.754
	1-2-gram	132	0.925	0.906	0.832	0.787	0.803	0.750
DAV	1-gram	9	0.791	0.736	0.750	0.682	0.678	0.592
	2-gram	81	0.911	0.888	0.797	0.742	0.783	0.727
	1-2-gram	90	0.919	0.898	0.792	0.737	0.777	0.719

Regarding classifier selection, SVM clearly outperforms RFs and NB for all datasets (see Tables 4.4 and 4.5 for a comparison).

Table 4.5.: Classification results for the Prot2Vec transformations, where N is the size of a feature set.

			SVM		RF		NB	
DATA	N	Accu	MCC	Accu	MCC	Accu	MCC	
Prot2Vec1	100	0.899	0.872	0.860	0.825	0.600	0.515	
Prot2Vec2	100	0.870	0.835	0.809	0.763	0.585	0.493	

Table 4.6 shows details, for the dataset of each alphabet with the best results, of the grid searches conducted to find the optimal parameters C and γ of the RBF-SVM: the combination of parameters C and γ found to have the best performance, and the corresponding mean accuracy and MCC values on the test sets are reported.

Table 4.7.: Subtype classification result obtained by SVM from the N -gram transformation of lengths 1-2 for the amino acid alphabet.

Subtype	MCC	Precision	Recall	Type I error	Type II error
mG	0.947	0.945	0.974	low	-
CS	0.934	0.958	0.918	low	low
GB	0.986	0.990	0.985	-	-
VN	0.911	0.936	0.927	low	low
Ph	0.880	0.894	0.931	medium	low
Od	0.774	0.880	0.706	medium	high
Ta	0.983	1.000	0.969	-	-

Table 4.6.: Model selection results for the dataset with the best results of each alphabet.

DATA	PARAMETERS	Accu	MCC
AA 1-2-gram	$C=[2,4]$, $\gamma=[2^{-9}, 2^{-10}]$	0.934	0.917
SEZ 1-2-gram	$C=[3,4]$, $\gamma=[2^{-7}, 2^{-8}]$	0.925	0.906
DAV 1-2-gram	$C=[3,4]$, $\gamma=2^{-6}$	0.919	0.898
Prot2Vec1	$C=[4,5]$ $\gamma=[2^{-6}, 2^{-7}]$	0.899	0.872
Prot2Vec2	$C=[3,5]$ $\gamma=[2^{-5}, 2^{-6}]$	0.870	0.835

The best classification results are found for the 1-2-gram transformed dataset from the AA alphabet using the SVM classifier achieving an accuracy of 0.934 and an MCC value of 0.917.

Table 4.7 shows the classification results for the 1-2-gram transformed dataset from the AA alphabet and the SVM classifier with greater detail at the per class level (these results correspond to a model with parameters $C=[2,4]$, $\gamma=[2^{-9}, 2^{-10}]$). The MCC value shows that classes mG, CS, GB and Ta are quite accurately discriminated from the other classes, having an MCC between 0.947 and 0.986. The prediction power of the classifier for classes VN, Ph and Od is clearly lower, with MCC values that range from 0.774 to 0.911.

As for the quality of the classifier, measured by the precision, it can be seen that it provides the most exact results for classes GB and Ta, as its precision gets very close to its maximum possible value. This metric shows that for classes mG, CS, VN, Ph and Od some type I classification errors (false positives) happen. Regarding the completeness of the classifier, measured by the recall, we see that it is most complete for classes mG, GB and Ta, which means that nearly all real

positives are correctly predicted. Classes CS, VN and Ph have a lower recall, meaning that some type II errors (false negatives) happen for these classes. Class Od has a significantly lower recall than the other classes, what means that this class is most difficult to recognize. Table 4.3 also shows an estimation of the quantity of type I and type II errors for each class.

4.1.4. Conclusion

The supervised, alignment-free classification with SVMs of the 2011 database version of class C GPCRs has been investigated in this experiments both using transformations based on the physicochemical properties of the AAs and n -gram representations from different alphabets. The experimental results using transformations based on the physicochemical properties have shown that the ACC transformed dataset has a clear advantage over the alternative transformations and that SVMs are best suited to the analysis of these data. Using n -gram representations the best results were found for the 1-2-gram transformed dataset and the AA alphabet, which clearly outperformed the other classifiers. Both the ACC and AA alphabet 1-2-gram achieved similar results with respectively an accuracy of 0.93 and 0.934 and an MCC value of 0.91 and 0.917.

The SVM classifiers built with this dataset and trained with the optimal parameters resulted highly accurate and discriminative. The per-class results have shown some differences regarding the prediction power for some subclasses, which encourage the analysis of the less distinctive classes and the related classification errors in a future work at the biochemistry level.

4.2. Experiments with the September 2016 dataset

4.2.1. Experimental settings

Methods In this Section, we turn to analyze the use of different alignment-free transformations of the complete sequence on the class C GPCR dataset published on September 2016 in GPCRdb, comprising 936 sequences. We analyze both transformations based on the physicochemical properties of the sequences, such as the ACC transformation, and a range of n -gram transformations, such as the n -grams of lengths one and two (AA and Digram respectively) from the AA alphabet and two variants of the Prot2Vec transformation. For a detailed description we refer the reader to the technical description in Section 3.3.2. The experiments are structured in three phases as for the 2011 dataset and use the same classification performance metrics for evaluation (See section 4.1.1) .

4.2.2. Results and Discussion

Table 4.8 shows the classification results obtained from different classifiers for the the Amino Acid Composition (AA), Digram Composition (Digram), Auto-cross covariance (ACC) and two variants of Prot2Vec: that based on a Swiss-Prot database representation and that based on a GPCRdb representation. The best results are obtained for the ACC transformation and SVM classifier, although the performance of the RF classifier is close to that of the SVM.

Table 4.9 shows details, for the datasets with the best results, of the grid searches conducted to find the optimal parameters C and γ of the RBF-SVM: the combination of parameters C and γ found to have the best performance, and the corresponding mean accuracy and MCC values on the test sets are reported.

Table 4.8.: Accuracy (Accu) and MCC according to dataset and classifier. N stands for the size of the feature set.

		SVM		RF		NB	
DATA	N	Accu	MCC	Accu	MCC	Accu	MCC
AA	20	0.989	0.982	0.985	0.976	0.959	0.937
DIGRAM	400	0.995	0.993	0.991	0.986	0.989	0.983
ACC	325	0.997	0.995	0.993	0.988	0.990	0.985
Prot2Vec1	100	0.989	0.984	0.986	0.978	0.990	0.985
Prot2Vec2	100	0.994	0.991	0.990	0.985	0.981	0.969

Table 4.9.: Model selection results.

DATA	PARAMETERS	Accu	MCC
AA	$C=3, \gamma=2^{-5}$	0.989	0.982
DIGRAM	$C=[4,6], \gamma=[2^{-13}, 2^{-12}]$	0.995	0.993
ACC	$C=6, \gamma=[2^{-11}, 2^{-10}]$	0.997	0.995
Prot2Vec1	$C=6, \gamma=[2^{-9}, 2^{-7}]$	0.989	0.984
Prot2Vec2	$C=[4,6], \gamma=[2^{-9}, 2^{-7}]$	0.994	0.991

The best classification results are found for the ACC transformed dataset using the SVM classifier achieving an accuracy of 0.997 and an MCC value of 0.995. Table 4.10 shows the classification results for the ACC transformed dataset and the SVM classifier with greater detail at the per-class level (these results correspond

Table 4.10.: Subtype classification result obtained by SVM from the ACC transformation of the 2016 version dataset.

Subtype	MCC	Precision	Recall	Type I error	Type II error
mG	0.998	0.996	1.0	-	-
CS	0.990	1.0	0.98	-	-
GB	0.994	1.0	0.990	-	-
Ta	0.998	0.996	1.0	-	-

to a model with parameters $C=6$, $\gamma=[2^{-11}, 2^{-10}]$. For all four subtypes, the classification is very accurate: the MCC ranges from 0.99 to 0.998. Considering also the measures of correctness (precision) and completeness (recall), the classifier achieves highly accurate results with values surpassing 0.996 and 0.98, respectively. For the 2016 version dataset, no recurrent classification errors are observed and no type I nor type II classification errors are found to exist.

4.2.3. Conclusion

The supervised, alignment-free classification with SVMs of the September 2016 version of class C GPCRs has been investigated through several experiments using a diverse range of transformations already used to analyze the older 2011 version dataset. The differences in performance between transformations for all classifiers are relatively small for the 2016 datasets and no transformation with no classifier falls below the 0.98 accuracy mark with the September 2016 dataset.

The experimental results have shown that the ACC transformed dataset achieves the best classification results. Regarding the comparison of classifiers, SVMs show an advantage over the other classifiers for all datasets, but also RFs and NBs obtain quite accurate results for the 2016 dataset. The SVM classifiers built with these dataset and trained with the optimal parameters resulted highly accurate and discriminative reaching an accuracy of 0.997 and MCC of 0.995. The subtype results have shown very accurate predictions for all subtypes and no recurrent misclassification errors have been observed.

5. A systematic approach to GPCR misclassification analysis

5.1. Introduction

Proteins have a rich taxonomy of families and subfamilies, for which the definition and use of class labels is necessary. The adscription of a protein to a family may be uncertain, or even wrong, thus becoming an instance of what has come to be known as a label noise (LN) problem. LN, which is commonplace in many scientific domains [94], has a potentially negative effect on any quantitative analysis of proteins that requires the use of label information. In fact, there are few domains in which the effects of LN are so pervasive as in biomedicine and bioinformatics [95]. The problem of LN may take many forms: from the human expert subjectivity in the labelling process, which is difficult to avoid, to bounds on the available information and communication noise [96].

In medicine, for instance, the reliability of diagnostic labels is often bounded by the natural limitations of the specialists' expertise [107], or even by the formal requirements of majority-based decision-making procedures, or consensus guidelines (for the latter see, for instance, [108]). In bioinformatics, protein subtype characterization is a task that is riddled with this problem, despite good practices in curation of genomic and proteomic databases [109].

In the specific field of GPCRs, which are the target of the current thesis, this problem is magnified by the fact that subtyping can be performed at up to seven levels of detail [97]. The occurrence of LN is unavoidable in this context because the assignment of individual sequences to one of these subtypes is itself, in most cases, a model-based process, which follows a complex many-step procedure that can only guarantee limited success [110].

In the research reported in previous chapters[3], we investigated the supervised classification of the class C GPCR dataset using different classifiers, namely RF, NB and SVM, for different alignment-free transformations of the sequences, including AA composition (AA) and the Mean Transformation and Auto-Cross Covariance (ACC). In this previous research, focus was placed on the accuracy of

the classifiers' performance and the experimental results showed that SVM clearly outperformed the rest of classifiers independently of the transformation applied to the data set. This led to the conclusion that a nonlinear classifier with the ability to find a linear separation of instances in a higher-dimensional feature space, such as SVM, was the adequate choice for the data set under analysis in the task of subtype discrimination. The second conclusion from this previous research was that, at subtype level, classification accuracies showed only small variations depending on data transformations. Even a superficial analysis of the confusion matrices showed recurrent patterns of subtype misclassification, which hinted at LN as their cause. Such observations provided support for a more detailed analysis of sequence misclassification.

In this chapter, we present a novel approach to assist the discovery of recurrent classification errors for supervised classification. The reported experiments using data from the curated GPCRdb database are meant to be the proof of concept for a novel systematic approach to assist the discovery of GPCR database labelling quality problems, which would in turn become the core of a label filtering decision support system [96], a useful tool for database curators in proteomics. Here, we analyze both the class C GPCR dataset from GPCRdb published in March 2011 and that published in September 2016.

The remainder of this chapter is structured as follows: The next section describes the novel approach to assist the discovery of recurrent classification errors followed by a report of the experimental results on the 2011 and 2016 version datasets and their discussion as well as the description of the experimental settings including the data transformations, methods of analysis and validation. The chapter wraps up with some conclusions.

5.2. Proof of concept for a novel systematic approach to assist the discovery of GPCR database labelling quality problems

In this section, we present a novel approach to assist the discovery of recurrent classification errors for supervised classification. First, we expose the systematic approach for misclassification analysis, which takes into account the metrics of the internal voting system and decision value of a SVM multi-class classifier.

The approach is exemplified using data from the curated GPCRdb database. We focus this investigation on the classification of data resulting from several alignment-free transformations of class C GPCR sequences into its several sub-

types. The sequences with the most consistent misclassification patterns are further analyzed to discover non-random LN effects, as a way to explore their possible biological explanation.

The candidate class C GPCR mislabelings detected using such approach are further validated through sequence visualization with phylogenetic trees (PT), dendrogram-like graphical representations of the evolutionary relationship between taxonomic groups which share a set of homologous sequence segments [111]. The visualization of the evolutionary relationship through PTs should serve as external validation tool and help in this study to confirm the correctness of the detected persistent mislabelings.

5.2.1. A systematic approach to misclassification analysis

The proposed method for the analysis of misclassifications comprises a method to evaluate SVM misclassification errors and an external validation approach with PTs.

Methods - A systematic approach to SVM misclassification analysis

Given a transformed dataset, our proposed systematic approach to the analysis of the classification errors consists of three steps or phases:

1. Estimation of the frequency of misclassification of each pattern (sequence) using different SVM models to select a subset of frequently misclassified patterns.
2. For each pattern in the subset selected in step 1, evaluation of the relation of votes of all the SVM classifiers between its true (label) class and its most-predicted class.
3. For each pattern in the subset selected in step 1, assessment of the decision values of the SVM binary classifiers between its true (label) class and its most-predicted class.

The aim of the first step is the detection of those patterns that, most of the times, are not classified as belonging to the class defined by their formal database label, but without considering the distribution of predicted classes in the misclassifications. Instead, the aim of the second and third steps is to confirm the consistency of the misclassifications to the most-predicted class. The difference between the two last steps resides on whether only the votes (i.e. the binary decisions of the SVM classifiers) are taken into account, or also the confidence (i.e. the decision values) of the binary SVM classifiers, when confronting just the class label against

the most-predicted class, are taken into account. The union of patterns obtained as a result in steps 2 and 3 forms the final subset of frequently and consistently misclassified sequences that are shortlisted as LN candidates. In the following subsections, further details of each one of the three steps are provided.

Repeated classification with different SVM models The first phase entails repeating the following procedure 100 times. Although this constant value could be changed, 100 is adequate both to obtain a statistically reliable result and to express the frequencies of misclassification directly as percentages (or error rates, ER_s , for each sequence s). This type of repeated CV approach has been proposed as well in [112] and applied in [113].

- First, the dataset is randomly reordered and a 5-fold CV (5-CV) is used, so that, for each of the five training-test partitions, the current training set is employed to construct an RBF-SVM model with an optimal value for the γ parameter of the kernel function and with the error penalty parameter C varying within a small range near its previously established optimum value.
- Second, a test set classification is carried out using the trained model, registering which GPCR sequences are misclassified and generating the corresponding confusion matrix.

The use of CV in each of the 100 repetitions of this procedure ensures that each instance is classified exactly one time as a test pattern in each iteration of the outer loop. Note that C is slightly modified in each iteration of the inner loop. With this, we obtain detailed results of how many times a sequence was misclassified when included in the test set and how many of these times it was assigned to specific classes. Note that all the classification results when the sequence belongs to the training set are not taken into account. In order to focus only on the most recurrent classification errors, a conservative misclassification boundary of $e = 75\%$ on the individual error rate ER_s was set (i.e., only sequences s misclassified in at least a 75% of the test occasions were deemed to be strong misclassifications and selected for further analysis). This threshold e is merely illustrative; in a real application of the method, it should be set according to the expert analyst's decision. A high threshold would ensure that only the most extreme misclassifications are singled out for further detailed analysis, whereas low thresholds would be more adequate in case a more global exploration is required.

Analysis of misclassifications according to the voting scheme Since we are facing a multi-class (K classes) classification problem in which the underlying classification scheme of the SVM implementation [59] was “one-*vs*-one”, it is interesting to analyze the results of the voting scheme as applied to the $K(K - 1)/2$

resulting classifiers, including the votes of each one, for each pattern in each test iteration. According to LIBSVM, the subtype with the highest number of votes in each case becomes the predicted class of the test pattern.

For each frequently misclassified sequence s , selected in the first phase, we focus the analysis on the relation between the total number of votes VT_s obtained by the true (label) class in the 100 iterations and those obtained by the most frequently predicted class for that sequence, VP_s . This is, we define the voting ratio

$$R_s = \frac{VT_s}{VP_s} \quad (5.2.1)$$

and, given some threshold θ_R , we consider that $R_s \leq \theta_R$ indicates a consistent (also deemed as *large*) classification error, while $R_s > \theta_R$ denotes a more doubtful (or *small*) misclassification. We fixed a threshold $\theta_R = 0.5$ to obtain our results discussed later.

Analysis of misclassifications according to the decision values In the third and last phase of our proposed approach, we go deeper into the analysis of misclassifications by taking into account the confidence (decision values) of the 100 binary SVM classifiers involving only the label class and the most frequently predicted class, when classifying a sequence s as test pattern. For each frequently misclassified sequence s selected in the first phase, we define a *cumulative decision value*, CDV_s , as follows:

$$CDV_s = \sum_{k=1}^{100} DV_s(i, j, k) \quad (5.2.2)$$

where $DV_s(i, j, k)$ is the decision value given by the binary SVM classifier confronting the class with label i to which s formally belongs and the most-frequently predicted class for sequence s , with label j , in the k^{th} test iteration. GPCR subtype labels were numbered 1 to 7 in the order they are presented in the data description section. For subtypes i, j , a large positive CDV_s value if $i > j$ and a large negative one if $i < j$ both indicate clear misclassifications. Hence, the magnitude of the error is deemed *large* or *small* depending on whether the CDV_s exceeds a certain threshold θ_{CDV} in absolute value or not. A threshold $\theta_{CDV} = 60$ was chosen for the experiments.

Note that the information conveyed by CDV_s complements that of R_s . For instance, a misclassified sequence with high R_s would suggest that the voting process

discards all subtypes but the true and the predicted ones, that is, a very narrow transfer of subtype assignment. If this is accompanied by a large CDV_s in absolute value, the predicted subtype, even if wrong assuming that the identifying label of the sequence is trusted, is strongly preferred by the SVM classifiers.

Methods - External validation of SVM-based classification

Mislabeling validation with phylogenetic trees Here, PTs are used to visualize the analyzed class C GPCR sequences and thus provide an alternative way to externally validate the misclassification results found with the proposed approach. There are two sound reasons why we use PTs for this task: first, because they have *de facto* become standard tools in bioinformatics [111] and, particularly, in protein homology detection, so that protein database curators are more likely to trust them. Second, because the protein sequence alignment that underlies the tree construction has no direct link with the sequence transformations from which the SVM classifiers are built, therefore guaranteeing the independence of the results.

Our software tool of choice, Treevolution¹ [42], was developed in Java and integrates the Processing² package. 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 at the user request and offers several methods very useful for large PT: sector distortion, tree rotation, pruning, labeling, tracking of ancestors and descendants and text search, among others.

The color-guided highlighting of protein families helps the user to focus on sequence groupings of interest and give an overall idea of groups with the same ancestor within the tree. The PT is created from a MSA obtained with Clustal Omega [114]. This application, in which sequences data are introduced in FASTA format, performs distance-based MSA [115].

5.3. Experiments on the class C GPCR dataset

In this section, we report the experimental results of the novel systematic approach for misclassification analysis using the 2011 and 2016 versions of class C GPCR dataset.

¹<http://vis.usal.es/treevolution>

²<http://processing.org>

5.3.1. Experimental Settings

Data preparation

In this study, six different transformations are used, where we distinguish between those based on the n-gram representation built on the AA alphabet and those based on the physicochemical properties of the AAs. Here, we use the AA and Digram methods, which transform the data according to the frequency of appearance of n-grams of, in turn, length one and length two in the sequence and the more sophisticated Prot2Vec distributed transformations. On the other hand, we decided to use more complex transformations based on the physicochemical properties of the AAs and the sequencing information such as ACC and PDBT.

Methods

We use the novel systematic approach to the analysis of the classification errors (as explained in section 5.2.1) in order to filter out the frequently and consistently misclassified sequences. First, the frequency of misclassification is estimated by using a repeated classification with different classification models. For the very frequently misclassified instances the related misclassification error is analyzed measuring the consistency of the classifiers decision by means of the voting ratio R_s and the cumulative decision value CDV_s . Finally the detected recurrent classification errors are validated externally with PTs as explained in section 5.2.1.

The classification results reported in this study are assessed by means of the accuracy and MCC for multiclass classification, whereas the subtype results use the MCC, precision and recall for binary classification. All experiments are conducted using 5-CV with stratified folds. SVM classification uses the LIBSVM library and an RBF kernel, whose parameters C and γ are again adapted by means of a grid search.

5.3.2. Mislabeling analysis of the 2011 dataset

This section describes the results of the analysis of LN on the class C GPCR dataset published in March 2011. In this experiment we use the AA, Digram, ACC and PDBT transformation.

Results

Repeated classification with different SVM models using different transformations of the dataset The previous results from [3, 4] led us to decide on the

Data	Accu	MCC
AA	0.88	0.84
Digram	0.93	0.91
ACC	0.93	0.91
PDBT	0.92	0.90

Subtype	MCC	Prec	Rec
mG	0.95	0.95	0.99
CS	0.93	1.00	0.88
GB	0.98	0.99	0.99
Vn	0.89	0.91	0.92
Ph	0.86	0.89	0.90
Od	0.79	0.89	0.74
Ta	0.99	1.00	0.98

Table 5.1.: SVM classifier results: Left: Global results for the four data transformations; accuracy (Accu), Matthews Correlation Coefficient (MCC); best results highlighted in bold. Right: Class C GPCR results per subtype for the ACC data set only, including MCC, Precision (Prec) and Recall (Rec).

convenience of using a more diverse set of data transformation techniques. Table 5.3.2 summarizes the best subtype classification results obtained with SVM for the four different transformed data sets. These results are complemented by the box-plot representation of the distributions of the accuracy and MCC values, for each of the transformed data sets, over the 100 outer iterations of the classification procedure, shown in Figures A.1.1 and A.1.2. For all transformations, a low variability of the results is observed, suggesting consistent estimates that make the average figures of Table 5.3.2 quite reliable. Out of these, the best classification results were found for the Digram and ACC transformed data sets, although the relative differences of accuracy and MCC make PDBT also a reasonable choice.

A detailed analysis of the results per-subtype revealed relatively minor differences between those obtained with each of the four transformed data sets. This observation suggests that the main causes of misclassification might lie beyond the differences between data transformations and that a more systematic analysis of the classification errors is required.

Table 5.2 shows a few illustrative misclassification statistics for the ACC transformed data set. For instance, sequence #6, which belongs to subtype *VN* according to its database label, was misclassified 100 out of 100 times: 96 of them was assigned to *Ph* and 4 to *Od* (See Table 5.3 for the mapping between the number # and the protein database *Id*).

This misclassification analysis was repeated for each of the transformed data sets. The number of considered misclassifications (those with $ER_s > 75\%$) was smaller for the ACC and Digram transformations - The AA, Digram, ACC and PDBT sets yielded, in turn, 143, 88, 85 and 100 strong misclassifications. A detailed analysis of these frequently misclassified sequences revealed that they are nearly identical

Table 5.2.: Illustrative example of misclassification statistics for the ACC data set. For some sequences s identified by number $\#_s$, the error rate (ER_s), the true class (TC_s), and how many times this sequence was misclassified as belonging to each of the other subtypes (from mG to Ta), are displayed. The three last columns list the sum of the votes for the true class (VT_s), for the most frequently predicted class (VP_s), and the ratio (R_s) of one to the other.

$\#_s$	ER_s	TC_s	mG	CS	GB	VN	Ph	Od	Ta	VT_s	VP_s	R_s
2	100	CS	100	0	0	0	0	0	0	91	600	0.15
6	100	VN	0	0	0	0	96	4	0	404	596	0.67
7	100	VN	100	0	0	0	0	0	0	300	600	0.5

for ACC and Digram. There are some differences with the PBDT misclassifications that might be the result of the very different type of transformation. Importantly, 52 frequently misclassified sequences were common to all four data sets and there was strong agreement on the most-often predicted subtypes. These sequences are listed in Tables A.1 and A.2.

Analysis of misclassifications according to the voting scheme Interestingly, these results suggest the existence of subtypes with recurrently wrong class assignments. So, we applied the second step of our systematic approach based on the voting scheme, as described earlier, to confirm consistent misclassifications. To illustrate the results obtained in this step, we show the voting scheme results for the selected instances of Table 5.2. Sequence $\#6$, for instance, is a *VN* consistently misclassified as *Ph*. The magnitude of the error is small, though, as the *voting ratio* (R_s) of true class to predicted class is relatively high ($0.67 > 0.5$). Sequence $\#2$ is a *CS*, consistently misclassified as *mG*. The magnitude of the error is large, as the R_s is quite low ($0.15 \leq 0.5$).

Only 7 of the 85 frequently misclassified ACC-transformed sequences yielded large errors (See Table 5.3). Similarly, for AA, Digram and PDBT sets, the majority of sequences have small errors.

Analysis of misclassifications according to the decision value Clear differences in the magnitude of the recurrent classification errors were found. Pursuing further insight, we applied the third step of our approach based on the *cumulative decision value* (CDV_s) specifically for the binary classifier that involves the true class and the predicted class.

As previously mentioned, the magnitude of the error was deemed *large* or *small* depending on whether the CDV exceeded the threshold of 60 in absolute value or

Table 5.3.: Sequences with large classification errors: For each sequence s numbered $\#_s$, the GPCRDB Identifier (Id_s), the true class (TC_s), the predicted class (PC_s), the *voting ratio* (R_s) and the *cumulative decision value* (CDV_s) are displayed. Extreme R_s and CDV_s values highlighted in bold.

$\#_s$	Id_s	TC_s	PC_s	R_s	CDV_s
1	q5i5c3_9tele	mG	Od	0.75	-95
2	XP_002123664	CS	mG	0.15	50
3	q8c0m6_mouse	CS	Ph	0.15	-46
4	XP_002740613	CS	mG	0	-66
5	XP_002936197	VN	Ph	0.83	-96
6	XP_002940476	VN	Ph	0.67	-95
7	XP_002941777	VN	mG	0.5	45
8	B0UYJ3_DANRE	Ph	mG	0.79	109
9	XP_001518611	Od	mG	0.31	46
10	XP_002940324	Od	VN	0.49	70
11	GPC6A_DANRE	Od	Ph	0.5	74

not. A total of 21 out of the 85 frequently misclassified instances of the ACC-transformed data set have a *large* error according to this criterion, whereof 4 yield a *very large* one ($|CDV_s| \geq 95$: see Table 5.3).

Summary of the analysis of misclassification The proposed subtype classification approach revealed the existence of a number of instances that, independently of the sequence transformation method, induce classification errors that could be deemed either large or small. The information provided by R_s and CDV_s should be understood as complementary, given that not fully coincident instances are singled out in each approach.

Importantly, this analysis showed that the misclassifications of a sizeable proportion of sequences have a small magnitude, so that they could be ignored unless a thorough revision of the database labels is required. A small number of instances, though, showed consistent and large classification errors and they should be the focus of interest from the database curation viewpoint. In Table 5.3, we list GPCRs with either very large absolute value of CDV_s (4 items) or small R_s (7 items) using the ACC transformed dataset.

Mislabeling validation

Validation through PT-based visualization of class C GPCRs Figure A.1.3 displays the *Treevolution* radial PT plot of the complete set of 1,510 GPCRs of class C, additionally showing the approximate distribution of its main seven subtypes. In this representation, each outer branch corresponds to one GPCR sequence. Tree colors are used to represent families of descendant nodes. Note though that these colors do not correspond to subtype labels. We observe that some families correspond to not one but several evolutionary branches. For example, the two different colors assigned to Pheromone provide quantitative evidence of the existence of at least two subtypes within the family. The representation of the evolutionary relationship in the PT plot shows that there exist some clearly separated families (GB, CS and Od), while others are more closely related to each other, such as mG and VN.

In the following, we report the PT plots for the four class C subtypes that were predicted for the mislabeling candidates listed in Table 5.3. In them, these potential mislabelings (those with largest errors according to the proposed approach) are highlighted (See the individual sequences listed in Table 5.3).

Figure A.1.4 shows the selection of sequences with largest errors that were predicted to be *mG*. The *mG* subtype has two main evolutionarily-related subgroups, which are shown schematically in the PT plot. In our analysis, we found 5 sequences with large classification error. In this PT, they are highlighted in their locations. We see that sequences #7 (labeled as *VN* in GPCRDB) and #2 (labeled as *CS*) both fall into the first area of *mG*. The instances #4 (labeled as *CS*), #8 (labeled as *Ph*) and #9 (labeled as *Od*) fall into the second area of *mG*.

Figure A.1.5 shows the single sequence with large error predicted as *Od* by the proposed mislabelling filtering approach. The Odorant subtype corresponds to a single area in the PT plot, and sequence #1 (labeled as *mG*) falls clearly into this area.

Figure A.1.6 shows the sequences found to be *Ph* by the proposed approach. The *Ph* subtype has two main evolutionarily-related subgroups, which are shown schematically in the PT plot. Sequences #11 (labeled as *Od*) and #3 (labeled as *CS*) fall into the first subgroup, whereas sequences #5 and #6 (labeled as *VN*) fall into a separate evolutionary branch.

Figure A.1.7 shows the sequence found to be *VN* by the proposed approach. The *VN* subtype corresponds to three evolutionary areas in the PT plot. Sequence #10 (labeled as *Od*) falls into one of these areas.

Comparison with an ensemble-based noise detection approach As mentioned in earlier sections, in previous studies we carried out classification experi-

ments on the AA and ACC datasets [3] using different supervised classifiers, including NB, RF and SVM. From these, we concluded that SVM classifiers significantly outperformed NB and RF for both the AA and ACC datasets. In this section, we return to these less accurate classifiers, which are more robust to LN as they apparently carry out a more generic classification of the investigated dataset, and should be less prone to data overfitting, a possible risk associated to the more accurate SVM classification models [116].

We describe here the results of the application of an ensemble-based noise detection approach including less accurate classifiers for the analysis of the class C GPCR data set. This method just detects label noise candidates by counting the misclassifications of an instance for the different classifiers in the ensemble. In some way, this is similar to the first step of our proposed approach, with the difference that in our case the ensemble is composed only by SVM classifiers. The second and third steps of the proposed approach, though, allow a more fine-grained analysis of the misclassifications by taking into account the results of the classifiers involving just the sequence class label and the most-frequently predicted class for that sequence. In fact, we do not aim to a straightforward comparison between methods, but to use the ensemble method as a way to test the coincidence on the subset of mislabeled sequences detected by both.

Ensemble-based noise detection methods have their origin in ensemble learning [54], where a set of prediction models are constructed using different algorithms and their output is combined to generate a single prediction. A noise detection ensemble classifier filter [100] consists of a set of diverse base classifiers. Their classification errors are combined to detect mislabeled instances using either a consensus vote filter (all classifiers detect a classification error), or a majority vote filter (the majority of classifiers detect an error).

In this experiment, we use an ensemble classifier built using NB, Random Forest (RF), SVM and Multi-Layer Perceptron (MLP) classifiers to analyze the ACC-transformed data set. The results of the base classifiers are evaluated using a noise rank filter [102], which provides information about the ranking of detected candidates to misclassification. The rank filter estimates the following weights (in brackets) for each classifier in the calculation of the rank: MLP (3), SVM (2), NB (1) and RF (1), and assigns a ranking to the sequences according to the number of classifiers that failed to evaluate them correctly. It then reports in how many classifiers the prediction failed. In our analyses, we focused on those sequences that were evaluated incorrectly by either all classifiers (a total of 117), or by at least three of them (a further 34). We then checked which of these 151 sequences were also detected as frequently misclassified by our proposed SVM approach. A total of 141 instances were found. Both methods coincided in the detection of 109 sequences as possible mislabelings (a 77% coincidence). All sequences with

large classification error listed in table 5.3 were also detected by the noise rank algorithm. This result provides further support to the claim of effectiveness of the proposed SVM-based approach in its task as LN detector.

Discussion

The systematic approach proposed for the analysis of the SVM misclassifications has revealed the existence of a number of sequences that, independently of the transformation method, are prone to classification errors that could be deemed large or small (according to criteria that, ultimately, should be set by proteomics experts). The information provided by the voting ratio R and the absolute value of the CDV should be understood as complementary, given that not fully coincident sequences are singled out by each approach; that is, some sequences might show very low values of R but not very high values of CDV, or very high values of CDV but not too low values of R .

Importantly, this analysis has shown that the misclassifications of a sizeable proportion of sequences have a small magnitude. All these sequences might well be considered as mild cases of LN and should eventually be redirected to a human expert for further analysis. Small errors also suggest underlying similarities between the GPCR subtypes whose characteristics may be unknown and worth investigating. A small number of instances, though, show consistent and large classification errors. They merit detailed study because they might be affected by a more radical type of LN, or even by straight mislabelling. These are the sequences listed in Table 5.3, which are now individually discussed.

Sequences *XP_002123664*, *XP_002740613*, *XP_002936197*, *XP_002940476* and *XP_002940324* are all recurrently misclassified. *XP_002740613*, in particular, yields a 100% error, $R = 0$ and large CDV. Their labels should require further expert assessment, given that they were derived by an automated computational analysis from an annotated genomic sequence by means of a gene prediction mode from the RefSeq³ databank. Another couple of interesting cases are *q8c0m6_mouse* and *B0UYJ3_DANRE*. According to the information referenced at UniProt⁴, these GPCRs are unreviewed and should be considered only as preliminary data. The former, according to GPCRDB, is a *CS* that our system confidently ($R = 0.15$) classifies as *Ph*. The European Nucleotide Archive⁵ lists it as similar to the putative *Ph* receptor V2R2. The latter, according to GPCRDB, is a *Ph*, while our system predicts it to be an *mG* with a very large CDV (109).

³<http://www.ncbi.nlm.nih.gov/refseq/>

⁴<http://www.uniprot.org/uniprot/{B0UYJ3,Q8COM6}>

⁵<http://www.ebi.ac.uk/ena/data/view/BAC26854>

Agreeing with our prediction, the Ensembl Genome Browser ⁶ considers it to be an *mG* of subtype *6a*.

Sequence *GPC6A_DANRE* is labeled as *Od*, but the low number of votes of this class and the large CDV suggest its classification to *Ph*. Although this sequence is considered as olfactory receptor⁷, we suggest to investigate the possibility of its labelling as *Ph*.

As stated in the previous section, it is important to provide further validation for the clearest of the misclassifications found with the proposed method (as summarized in Table 5.3) using PTs. The importance of this validation resides in the fact that the PT dendrograms are not built from the same data transformations we used. Therefore, agreement between the subtype assignment of the PT and the label predicted by our method should be an almost definitive confirmation of the existence of label noise, whereas, contrarily, lack of agreement might be an indication that the misclassification is caused by the type of sequence transformation itself, or by the fact that the subtypes defined by the existing and predicted labels overlap.

The comparison of the most extreme misclassifications discovered with the proposed method with the visual results provided by the PTs (See Figures A.1.4 - A.1.7) is striking, as it provides consistent evidence of the reliability of the former. Figures A.1.4 - A.1.7 show that the detected extreme mislabellings fall exactly into the evolutionary branch belonging to the class predicted by the proposed approach. This reliability is a guarantee that the method is viable as a tool for database curators in proteomics.

Conclusions

LN is a potentially important problem in the process of automated class C GPCR subtype classification from the alignment-free transformed versions of protein primary sequences. This is because the labels of these sequences are obtained indirectly through complex, many-step similarity modelling processes.

In this study, we have proposed a systematic procedure, based on SVM classification, to single out and characterize GPCR sequences with consistent misclassification behaviour. This approach, where the detection of possible mislabeled data is based on the analysis of the frequency of misclassification of an instance and a quantitative assessment of the magnitude of the classification error, has been

⁶<http://www.ensembl.org>

⁷<http://www.uniprot.org/uniprot/Q5U9XR>

applied to different sequence data transformations and shown to be a viable alternative for the definition of a prediction-based system addressing the problem of LN.

For a database like the one analyzed in the current study, the type of LN is well-defined within the general taxonomy of the problem [94]: it should not be mistaken by a problem of outlier or anomaly detection and can be considered as the natural result of human expert involvement and model-based (semi)automated labeling [117]. As such, it falls within the *noisy not at random* type of models, because sequences are more likely to be mislabeled when they are similar to sequences of other subtype and because labels are likely to be less certain in regions of low data density. Mislabeleding thus depends both on the data features and on the true labels. Three general (and partially overlapping) approaches are available to tackle this problem: the use of classification algorithms that are robust to LN; the use of *filter* methods that detect *noisy* cases; and the use of algorithms for explicit LN modeling. A large palette of methods has been proposed for each of these and their review is beyond the scope of this study. The reader is referred to [94] for an up-to-date survey.

Here, our choice was a variant of a filtering approach, because, as acknowledged in [94], “some of the label noise-tolerant variants of SVMs could also be observed as filtering”. The proposed method can therefore be considered as model predictions-based filtering [118], extending the basic concept of voting filtering [100, 119] and attempting to improve model robustness by decomposing a multi-class problem into multiple binary classification problems [120]. The reported experimental results are a proof of concept for the viability of such procedure as part of a decision support system that, combined with expert knowledge in the field, should be able to assist the discovery of GPCR database labelling quality problems. These results have been further validated using PTs, a standard tool in bioinformatics.

5.3.3. Mislabeleding analysis of the 2016 dataset

Results

Repeated classification with different SVM models using different transformations of the dataset In this section we detail the experimental results of the analyses of the class C GPCR dataset published in September 2016 comprising 936 sequences. We report the classification results obtained using SVM classifiers for the transformed primary sequences of the proteins applying 5-CV: AA, Digram, ACC, and two variants of Prot2Vec: the first based on a Swiss-Prot database representation and the second based on a GPCRdb representation.

Data	Accu	MCC
AA	0.989	0.982
Digram	0.995	0.976
ACC	0.997	0.995
Prot2Vec1	0.989	0.984
Prot2vec2	0.994	0.991

Subtype	MCC	Prec	Rec
mG	0.998	0.996	1.0
CS	0.990	1.0	0.980
GB	0.994	1.0	0.990
Ta	0.998	0.996	1.0

Table 5.4.: SVM classifier results: Left: Global results for the four data transformations; accuracy (Accu), Matthews Correlation Coefficient (MCC); best results highlighted in bold. Right: Class C GPCR results per subtype for the ACC data set only, including MCC, Precision (Prec) and Recall (Rec).

Model	TC	PC	ER_s	R_s	CDV_s	TC	PC	ER_s	R_s	CDV_s
AA	GB	Ta	100	0.49	38.18	mG	Ta	100	0.34	-59.58
Digram	GB	Ta	96	0.51	-9	mG	Ta	100	0.38	28.75
ACC	GB	mG	100	0.46	19.16	mG	mG	0	-	-
Prot2Vec1	GB	Ta	100	0.58	-42.54	mG	CS	100	0.33	55.5
Prot2Vec2	GB	Ta	100	0.41	-28.52	mG	CS	100	0.39	-10.36

Table 5.5.: Analysis of misclassification of sequences *h2u5u4_takru* and *t2mdm0_hydvu*: For each sequence s the true class (TC), the predicted class (PC), the error rate (ER_s), the voting ratio (R_s) and the cumulative decision value (CDV_s) are reported.

Summary of the analysis of misclassification A study of the misclassifications of the 2016 database reveals that only the sequence *h2u5u4_takru*, labeled as GB, is misclassified for all the data transformations of the present study. Nevertheless, the prediction of class membership of this sequence is not completely uniform, as it is predicted to belong to Ta in four cases and to mG in one case. This is, according to Uniprot, an uncharacterized protein, i.e. inferred from homology. Sequence *t2mdm0_hydvu* was also detected as frequently misclassified (for 4 out of 5 transformations). This sequence is labeled as mG, but the classifiers predict it to belong to CS or Ta. Table 5.5 details the measures employed to analyze the consistency of misclassification of these two sequences.

Discussion

For the September 2016 version dataset comprising 936 sequences the classification model achieved an accuracy of 0.997 what implies only three misclassifications. As

reported in the previous section, just two sequences were very consistently misclassified and might be considered as possible cases of LN: *h2u5u4_takru*, labeled as GB and predicted to probably belong to Ta and *t2mdm0_hydvu*, labeled as mG and predicted to belong to CS or Ta. In contrast, the results from the study of the 2011 database [5], revealed the existence of at least 11 consistent misclassification of sequences using the same conservative thresholds to assess the consistency of data labeling.

Conclusions

The experiments on the 2016 version database indicate that this more recent version of the class C GPCR dataset is very accurately labeled and almost free of the LN problem as the consistently misclassified sequences detected by the proposed ML-based approach are now very scarce. This means, that although misclassifications may happen, the method considers that these cases should not be seen as labeling errors.

6. Tracking the evolution of class C GPCR database using machine learning tools for biocuration assistance

6.1. Introduction

Biocuration in biology in general and specifically in the omics sciences has become paramount, as research in these fields swiftly evolves towards an increasingly data dependent model. As a result, the maintenance and management of web-accessible publicly-available databases becomes a central task in biological knowledge dissemination. One relevant challenge for biocurators is the unambiguous identification of biological entities. In this thesis we analyze class C GPCRs. These receptors are characterized according to subtype labels at different levels of organization. The research reported in the previous chapter has provided evidence that some of these receptors could be affected by a case of LN, as they appeared to be too consistently misclassified by ML methods. Here, we review the evolution of class C GPCR database from the former 2011 version to two recent and quite substantially modified new versions of this database from 2016 assessing the internal data quality using supervised classification methods and focusing on possible mislabellings through a LN analysis. The analysis reveals the now extremely accurate labeling for the new versions of 2016 using several ML models and different transformations of the unaligned sequences. These findings support the adequacy of our proposed method to identify problematic labeling cases as a tool for database biocuration.

6.2. Comparison of class C GPCR datasets

6.2.1. Experimental settings

Methods GPCRdb [38] is a web-accessible and publicly-available repository and information system containing data and web tools especially designed for GPCR research. Class C, investigated in the current study, in turn comprises several subtypes. Receptor databases are regularly updated. A class C dataset from March 2011 was object of extensive analysis using ML methods in previous research [3, 5, 12, 121, 72].

In the current study, we go one step further and track the evolution of the class C GPCR dataset from GPCRdb comparing the 2011 dataset with two recent and successive versions from 2016 (May and September). We analyze the following aspects of the datasets:

1. We compare the datasets regarding the number of sequences and the number of subtypes of the class C GPCRs.
2. As an assessment of the internal data quality of the datasets we use different ML techniques. More specifically, we first compare the classification performance of several supervised classifiers, namely RF, NB and SVM for the different transformed datasets in order to decide which classifier is most adequate for the rest of the analysis. In this study we use several data transformations: the Amino Acid Composition (AA), Digram Composition (Digram), Auto-cross covariance (ACC) and two variants of Prot2Vec: the first based on a Swiss-Prot database representation and the second based on a GPCRdb representation. The classification performance is measured by means of the accuracy (Accu) and Mathew's correlation (MCC) coefficient for multiclass classification. The subtype results are evaluated by means of the MCC, precision and recall for binary classification. All experiments are conducted using 5-CV with stratified folds. SVM classification uses the LIB-SVM library and a RBF kernel, whose parameters C and γ are adapted by means of a grid search.
3. As an assesment of data labeling issues we compare the results of the analysis of frequently misclassified items using the 'systematic approach for mislabelling analysis', which was described in detail in section 5.2.1 with the respective results for the class C GPCR datasets.

6.2.2. Experimental results

Context Biological information, mainly in the omics sciences, is usually curated by specially assigned professional scientists in a process often known as biocuration. It has been described as “the activity of organizing, representing and making biological information accessible” [122] to biologists. It is becoming a key task, given that expert-curated web-accessible databases are one of the main driving forces in current biology in general and bioinformatics in particular [123]. The responsibilities of curators may include data collection, consistency and accuracy control, annotation using widely accepted nomenclatures, evaluation of computational analysis, etc. Biocuration requires broad expertise in the domain because of the vast amount of heterogeneous information available from literature, often lacking a unified and standardized approach for the representation and analysis of data. This often involves a previously unforeseen forefront role for text mining methods [124]. One of the challenges of biocuration is the unambiguous identification of biological entities (the class C GPCRs in this study) from existing studies and literature. Data trustworthiness can only be ensured through costly data management [125]. This task, when understood as “manual” expert curation, is uncertain and error-prone, so that the development of computational procedures to assist experts in it is worth pursuing.

Prior analyses of the 2011 database version revealed a possible sequence LN problem (described in more detail in the next section) in the form of sequences clearly and consistently predicted by classifiers as belonging to a different subtype than that reflected by their database label. These analyses were understood as laying the foundations for the development of a tool to assist database experts (not restricted to GPCRdb) in their curator tasks by shortlisting proteins with questionable subtype labels.

In short, the main motivation of the current study is to track the evolution of the GPCRdb database from 2011 to 2016 using the class C primary sequence data in order to find out whether the LN problem might have been successfully tackled, ameliorating classification. If so, this would reinforce the validity of our LN analysis methodology as a tool for biocuration assistance.

Data

The GPCRdb [37, 38] is a curated and publicly accessible repository of GPCR databases and web tools for the analysis of membrane proteins including about 400 human specimens. Overall, the GPCRdb dataset contains 14,951 proteins from 3,184 species.

The categorization of the receptors available from this database follows the international IUPHAR system recommendations. The whole database originally consisted of seven families: A (Rhodopsin), B1 (Secretin), B2 (Adhesion), C (Glutamate), F (Frizzled), Taste 2 and “other” GPCRs. This classification followed the system suggested in [40]. As mentioned in the introduction, the computational experiments reported in this study concern GPCRs of class C. At the highest level of grouping, class C discriminates receptors as *ion*, *amino acid*, or *sensory* according to the type of ligand.

Evolution of the database This study covers the evolution of GPCRdb over three versions: the first one released in 2011 and two recent and drastically changed versions: those of May 2016 and September 2016. At the second level of classification of the current database version, four subtypes are distinguished: metabotropic glutamate receptors (mG, amino acid), GABA_B (GB, amino acid), calcium sensing (CS, ion) and taste 1 receptors (Ta, sensory), covering sweet and umami tastes. The earlier 2011 version of the database also included three more sensory-related subtypes of the second level, namely vomeronasal (VN), pheromones (Ph) and odorant (Od) receptors.

Over the five years elapsed between the earlier and later versions of the database analyzed in this study, GPCRdb has undergone major changes in the total numbers of proteins belonging to class C, but also in the ratio of the different subtypes to the total number of receptors and even in the sequences contained in each of those subtypes.

Mere comparison of the datasets shows a remarkable reduction of the number of sequences, from the 1,657 sequences of the March 2011 dataset, down to the 954 of the September 2016 one, including orphans. Moreover, the variety of subtypes included in class C has been reduced from the seven of the 2011 dataset to only four in both 2016 datasets. (see Table 6.1 and Figure 6.2.1 for some summary figures).

The main changes occurred in the transition from the 2011 to the May 2016 versions, with only 155 protein sequences remaining unchanged. Not only the receptors of three subtypes (VN, Od and Ph) were removed in full from class C, but the number of proteins in the other remaining subtypes also changed significantly.

The mG receptors subtype grew by 33% and only 26% of sequences were kept unchanged (2011 \cap May 2016 column in table 2.1). The CS receptors subtype more than doubled, keeping only 10 sequences unchanged. Finally, the Taste 1 subtype grew threefold (note that in the the 2011 version it was characterized simply as Taste), while the GB receptors subtype, on the contrary, decreased more than threefold.

Subtype	March 2011	May 2016	Sept 2016	2011 \cap May 2016	May 2016 \cap Sept 2016
mG	351	467	516	93 (26%)	357 (76%)
CS	48	125	103	10 (21%)	91 (73%)
GB	208	60	89	10 (5%)	50 (83%)
Ta	65	193	228	42 (65%)	129 (67%)
Vn	344	0	0		
Ph	392	0	0		
Od	102	0	0		
Orphans	147	193	18	0	18 (9%)
Total	1657	1038	954	155	645

Table 6.1.: Number of receptors in each subtype for the class C datasets of the different database version, including percentages of sequences preserved from one version to the next.

The changes between the two 2016 versions are not so substantial, but still significant for a mere four-month period. In this case, the number of sequences kept completely unchanged varied from 65% to 85% for the four subtypes. The mG subclass kept growing in the September 2016 version by 10%; the GB and Taste 1 also increased by 50% and 18%, respectively. Instead, CS decreased by 18%. The largest of differences, though, was to be found in the number of orphan receptors (those not assigned to a subclass). Less than 10% of the original orphans were kept in the last version.

Previous research on GPCR class C from a data curation perspective

Subtype classification of GPCRs has been attempted at different levels of detail [97]. Our interest in the analysis of the evolution of this database from a data curation perspective stems from early experiments [3] in which we tested the extent to which class C GPCR first-level subtypes could be automatically discriminated from different transformations of their unaligned primary sequences.

Work on the 2011 version of the database provided evidence of clearly defined limits to the separability of the different class C subtypes. This evidence was produced using both supervised [12, 121] and semi-supervised [72] ML approaches and from different data transformation strategies. Interestingly, the subtypes shown to be most responsible for such lack of complete subtype separability were precisely those which were removed in the 2016 versions of the database (namely vomeronasal, odorant and pheromone receptors).

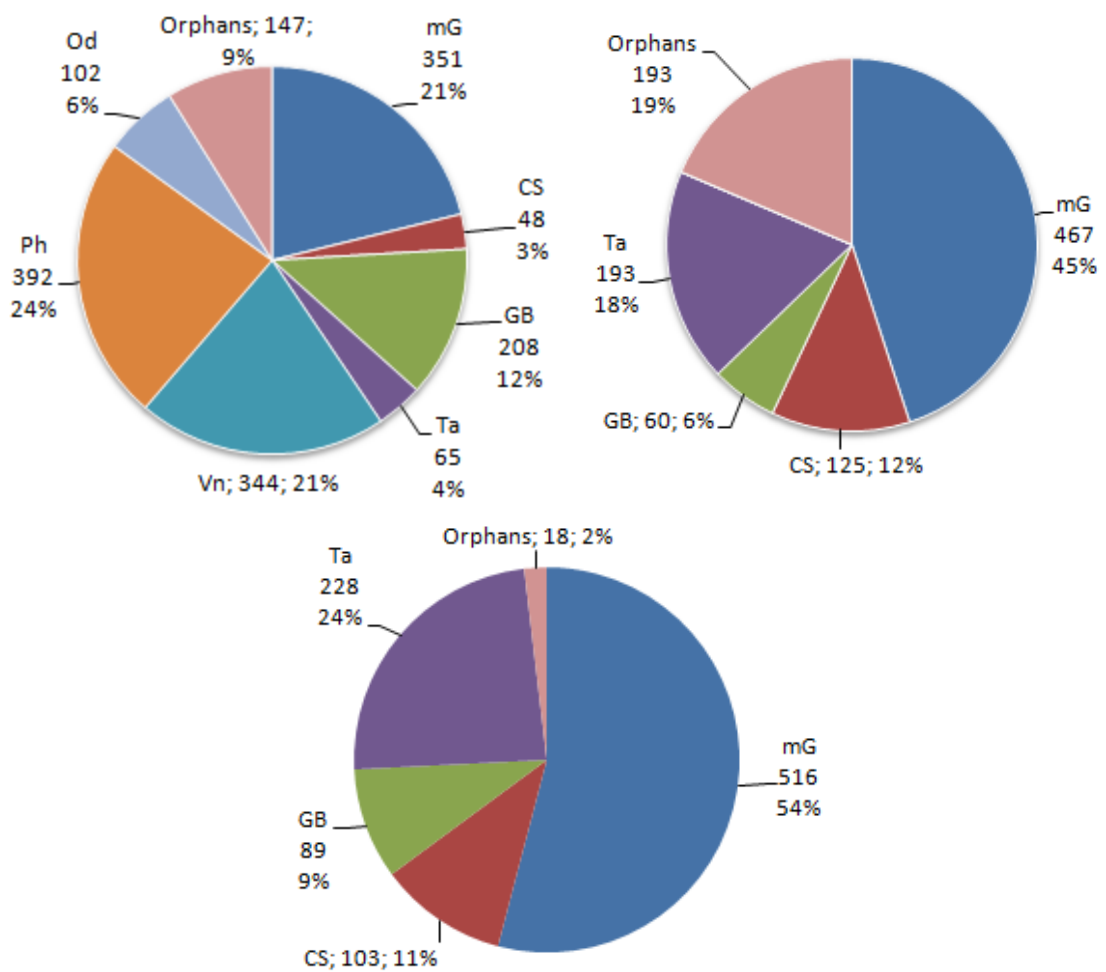


Figure 6.2.1.: Subtype distribution for the different databases (with orphans): upper left - March 2011, upper right - May 2016, middle - September 2016

These results were further confirmed from the viewpoint of visualization-oriented fully unsupervised machine learning methods (that is, methods that attempted sequence discrimination without knowledge of sequence-to-subfamily adscription). Results clearly indicated that the subtypes shown to be worse discriminated by supervised classifiers were also those shown to heavily overlap in unsupervised visualization models from different unaligned sequence data transformations [6].

These results might be just considered as a typical case of heterogeneous levels of class separability, often observed in real biological datasets. Closer inspection of the sequence misclassification behavior, though, revealed an intriguing and potentially more interesting pattern: different training runs of the same, or even of different, classification algorithms, might be expected to yield different class pre-

dictions for the same sequences. That is, we might expect a given sequence to be misclassified only in part of the experiments and/or be misclassified to different classes (subtypes in this case). For instance, a receptor sequence might be misclassified in only a percentage of experiments, being perhaps sometimes predicted to be a CS receptor, while others predicted to be a GB receptor. Some of the observed misclassifications conformed to this typical pattern, but many others were found to be far too consistent, in the sense that the sequence was almost always misclassified (by different classifiers and different implementations of the same classifier) as belonging to the same *wrong* subtype.

This behaviour suggested that we might be witnessing a case of the LN problem [94]. This is, the possibility that the sequence subtypes labels as appearing in the database, taken to be the ground truth, were actually wrong as the result of the uncertainty of the own database sequence labeling procedure, very often model-based itself. This would explain both the presence of consistently misclassified proteins and (at least partially) the limits of subtype discrimination accuracy which our experiments stubbornly showed to exist, independently of the choice of data transformation and classification technique.

This problem was analyzed in detail in [5], where individual sequences were identified and shortlisted as potential cases of LN to be further analyzed by database curators. Unsurprisingly perhaps, most of them belonged to the same three subtypes previously identified as the most difficult to discriminate, namely Vn, Ph and Od. All data transformations used in these experiments were alignment-free and included n -gram frequencies for $n = 1, 2$, auto-cross-covariance (ACC) and the physicochemical distance transformation (PDBT). The classifier of choice was a SVM, a model that has been widely favoured for this type of problems (see, for instance, [126, 48, 47]).

Subsequent work reported in [121], which again employed alignment-free data transformations, used a RF classifier to further investigate the consistency of misclassification in this problem. Note that RF is an ensemble learning technique with an internal classification *voting* system that is naturally suited to classification consistency analyses. The classification performance achieved with RF was similar to that of SVM across transformations. Most consistent misclassifications were again detected mainly in Vn, Ph and Od, confirming our previous results.

All these studies were based on the earlier 2011 version of the database, which automatically raised the following research question: if the 2011 database, which included Vn, Ph and Od as Class C GPCR subtypes, suffered from these LN classification problems, would the new 2016 versions of the database, which do not include those subtypes, suffer from similar problems? This is the question we aim to answer through the experiments reported next.

Results

In this section we detail the experimental results of the analyses of the three different datasets. We report the classification results obtained using different supervised classifiers for the transformed primary sequences of the proteins applying 5-CV. Tables 6.2 and 6.3 show, in turn, the classification results for the datasets published in March 2011, May 2016 and September 2016. In each table, several evaluation measures (Accu, MCC and F-measure) are reported for SVM, NB and RF classifiers, as well as for five different transformations of the primary sequences: the Amino Acid Composition (AA), Digram Composition (Digram), Auto-cross covariance (ACC) and two variants of Prot2Vec: the first based on a Swiss-Prot database representation and the second based on a GPCRdb representation.

Model	Classifier	Accu	MCC	F-meas
AA	SVM	0.8855	0.8549	0.8842
	RF	0.8570	0.8207	0.8542
	NB	0.7033	0.6307	0.7046
Digram	SVM	0.9311	0.9128	0.9303
	RF	0.9139	0.8929	0.9124
	NB	0.8358	0.7949	0.8375
ACC	SVM	0.9252	0.9054	0.9234
	RF	0.8894	0.8624	0.8838
	NB	0.8430	0.8064	0.8455
Prot2Vec1	SVM	0.8987	0.8715	0.8981
	RF	0.8596	0.8245	0.8587
	NB	0.6000	0.5153	0.6070
Prot2Vec2	SVM	0.8695	0.8353	0.8692
	RF	0.8093	0.7625	0.8110
	NB	0.5854	0.4931	0.5889

Table 6.2.: Classification results for the 2011 version dataset. *Prot2Vec1* corresponds to the Swiss-Prot-based representation and *Prot2Vec2* corresponds to the GPCRdb-based representation.

The best classification results were obtained with the SVM classifier for all three datasets. Tables 6.4 and 6.5 detail the classification results at the subtype level.

A detailed analysis of the consistently misclassified sequences reveals no coincidence with the results from the study of the 2011 database [5], for the obvious reason that none of the 11 sequences reported as consistently misclassified in this study (See Table 5.3) is part of the 2016 databases (for a formal description of the misclassification consistency concept, we refer the readers to the description

Model	Classifier	May 2016			September 2016		
		Accu	MCC	F-meas	Accu	MCC	F-meas
AA	SVM	0.9822	0.9714	0.982	0.9893	0.9824	0.9892
	RF	0.9716	0.9538	0.9706	0.9850	0.9757	0.9850
	NB	0.9550	0.9271	0.9551	0.9594	0.9368	0.9598
Digram	SVM	0.9917	0.9884	0.9916	0.9946	0.9925	0.9946
	RF	0.9905	0.9847	0.9905	0.9914	0.9860	0.9914
	NB	0.9811	0.9688	0.9808	0.9893	0.9826	0.9893
ACC	SVM	0.9941	0.9917	0.994	0.9968	0.9951	0.9968
	RF	0.9893	0.9830	0.9891	0.9925	0.9878	0.9925
	NB	0.9799	0.9673	0.9798	0.9904	0.9845	0.9903
Prot2Vec1	SVM	0.9822	0.9716	0.9822	0.9893	0.9839	0.9893
	RF	0.9763	0.9612	0.9759	0.9861	0.9776	0.9861
	NB	0.8118	0.7229	0.8207	0.9904	0.9845	0.9903
Prot2Vec2	SVM	0.9822	0.9759	0.9823	0.9936	0.9912	0.9936
	RF	0.9822	0.9714	0.9821	0.9904	0.9847	0.9903
	NB	0.8615	0.7972	0.8688	0.9808	0.9692	0.9809

Table 6.3.: Classification results for the May and September 2016 version datasets respectively.

Subtype	Precision	Recall	MCC	F-measure
mG	0.9462	0.9829	0.9639	0.9532
CS	1.0	0.9356	0.9645	0.9652
GB	0.9905	0.9856	0.9880	0.9861
Vn	0.9185	0.9128	0.9153	0.8907
Ph	0.8980	0.9131	0.9050	0.8719
Od	0.8610	0.7362	0.7896	0.7806
Ta	1.0	0.9846	0.9920	0.9918

Table 6.4.: Subtype classification results obtained by SVM from the Digram transformation of the 2011 version dataset.

about LN in section 5.2.1). A study of the misclassifications of the 2016 database reveals that only the two sequences appear as frequently misclassified in this more recent version of the database (see section 5.3.3 for the detailed results).

Subtype	Precision	Recall	MCC	F-meas	Precision	Recall	MCC	F-meas
mG	0.9958	1.0	0.9979	0.9953	0.9962	1.0	0.9981	0.9957
CS	0.9923	0.9760	0.9833	0.9811	1.0	0.9804	0.9899	0.9889
GB	1.0	0.9833	0.9913	0.9909	1.0	0.9889	0.9943	0.9938
Ta	0.9903	0.9949	0.9924	0.9902	0.9957	1.0	0.9979	0.9972

Table 6.5.: Subtype classification results obtained by SVM from the ACC transformation of the May and Sept. 2016 version dataset respectively.

Discussion and Conclusions

Note that the main goal of this study is the comparative analysis of class C GPCR data over time using three versions of a publicly available database spanning from 2011 to 2016. This analysis concerns the ability of different ML methods to discriminate between class C subtypes from different transformations of their unaligned sequences. Such discriminability analysis is geared towards the assessment of the LN problem observed in our previous investigation of the 2011 version datasets and is meant as a way to assist database experts in their biocuration tasks.

The mere comparison of the datasets shows a remarkable reduction of the number of sequences, from the 1,510 sequences in the March 2011 dataset, down to the 936 collected in the September 2016 one, not counting orphans. Moreover and as previously mentioned, the variety of subtypes included in class C has been reduced from the seven of the 2011 dataset to only four in both 2016 datasets.

The results of the analyses of the datasets using supervised classification methods, reported in the previous section, lead to some unequivocal conclusions.

According to the results in Tables 6.2 and 6.3, all classifiers perform better with the 2016 datasets than with the 2011 dataset according to all the evaluation measures considered. Furthermore, the September 2016 version of the dataset yields consistently better results than the May 2016 version although, in this case, differences are comparatively minor.

It might be argued that the differences between the 2011 and 2016 datasets could be put down to the fact that the Vn, Ph and Od subtypes have been removed from the 2016 versions. This is true only to some extent because, importantly, the subtype-specific results in Tables 6.4 and 6.5 indicate that the 2016 versions yield better performance than the 2011 version for each and every of the four remaining subtypes independently. And again, the September 2016 results are slightly better than the May 2016 results for each of the four subtypes.

An accuracy of 0.9941 using the SVM with ACC transformation for the 845 sequences of the May 2016 version dataset implies just 6 misclassifications. Correspondingly, a 0.9968 accuracy, also for the SVM with ACC for the 936 sequences of the September 2016 version dataset, implies 3 misclassifications. These are almost negligible numbers when compared to those of the 2011 version. Moreover, note that out of these few cases and as reported in section 5.3.3, only a couple of sequences show the type of very consistent misclassification that might be evidence of LN. In comparison, the results from the study of the 2011 database [5, 4], using the same criteria as the current study, indicated the existence of a shortlist of at least 11 very consistently misclassified sequences even when an extremely conservative threshold was used to assess such consistency. In our opinion, this is evidence of sound curation at work, as well as evidence of how important it is to use LN detailed assessment as a tool to assist biocuration.

We can also conclude that SVM classifiers show a very consistent overall advantage when compared to RF and NB for all three datasets and for all five data transformations. The difference is very clear with the 2011 version and more nuanced with the 2016 datasets. This is a relevant result for two reasons: first, because it reveals SVM performance to be more robust in datasets with limited class separability; second, because it reveals that with neatly separable classes such as those of the 2016 datasets, almost any classifier will do reasonably well, even the baseline NB classifier. This is further evidence that sound biocuration, when dealing with the LN problem adequately, helps to reduce the uncertainty associated to model-based decision making, in this case by limiting the impact of the choice of data analysis methods (here, the choice of classifiers) on the results.

Finally, we should consider the impact of the data transformations on the classification results. The interpretation of the corresponding comparative results bears similarities with that of the comparative of classification methods. Digram performs best for the 2011 version of the database, while the more complex ACC performs best for both 2016 versions. Again, the differences in performance between transformations for all classifiers are relatively small for the 2016 datasets and no transformation with no classifier falls below the 0.98 accuracy mark with the September 2016 dataset. Therefore, this again reinforces the idea that biocuration, by dealing with LN, reduces the uncertainty associated to model-based decision making, in this case by limiting the impact of the choice of data transformation method on the results. A last comment on this issue is that the recently proposed (and most complex of our choices in this study) Prot2Vec transformation [78] does not seem to show any relative advantage for the analyzed data.

Our experiments quite conclusively indicate that the last 2016 version of the class C data in GPCRdb, a reference for GPCR research, is almost free of the LN problem. That is, almost none of the class C GPCR sequences in this version

is predicted by our ML-based method to be consistently misclassified. In other words, the method considers that, even if misclassifications still exist, almost none of them should be suspected to be a labelling error. Having tracked this database from 2011 according to this criterion, we are now in a position to confidently say that the analysis of label noise in this type of databases, understood as a problem of misclassification consistency, is a useful tool for biological database curation. Importantly, and despite the fact that the research reported in this paper has focused on class C GPCRs due to their particular pharmacological relevance, the proposed method could be *exported* to any database in which biological entities are associated to a characterization label. This research also highlights the importance of documenting the reasons for changes between versions of publicly available biological databases.

7. Topological sequence segments discriminate between class C GPCR subtypes

7.1. Introduction

In this chapter, we systematically analyze whether segments of the receptor sequences are able to discriminate between the different class C GPCR subtypes according to their topological location on the extracellular, transmembrane or intracellular domain.

In the research described in previous chapters, we investigated the feasibility of discrimination between the defined subtypes of class C GPCRs using supervised ML classification approaches. As basis of the construction of the classification models, different alignment-free sequence transformations were used, including both transformations based on the physicochemical properties of the AAs [3] and on short n -gram features [10]. These experiments showed clear differentiation between the subtypes, but also an evident upper threshold to classification accuracy as well as some consistent misclassification patterns [5, 4]. Note that these former experiments were based on the entire and unaligned primary sequences of the receptors.

The GPCRs have different structural domains, including, amongst others, a seven-helix transmembrane (7TM) domain and an extracellular domain. In the case of class C, they include a large domain in the extracellular part of the receptor (N-terminus), which is built by the Venus Flytrap (VFT) and a cystein rich domain (CRD) connecting both in many of their subtypes [26].

Here, we now provide a systematic analysis of the subtype discrimination capabilities of the complete set of different topological locations in the class C sequences (in extracellular, transmembrane and intracellular domains), including their combinations. We compare this with the performance of the models based on the complete sequence.

We analyze to what degree the different topological parts of a GPCR retain the ability to discriminate between subtypes of the complete sequence, as analyzed in

previous research.

7.2. Experiments on the class C GPCR dataset

Class C GPCR subtype discrimination is addressed here as a supervised classification problem in which class labels are the assignments of each of the sequences to one of the existing subtypes according to the database available information. We measure in which degree the class C GPCR segments discriminate properly between the subtypes. We analyze both the class C GPCR dataset from 2011 and the September 2016 dataset.

7.2.1. Experimental Settings

Methods

The first phase of the experiments reported in this chapter involve the use of several models for the classification of the alignment-free complete sequences. These results were used to choose which classifier was most adequate for the rest of analyses. The comparison was performed with a similar selection of classifiers to those used in a previous study [3], which include NB, RF and SVM classifiers.

The classification results for all classifiers were achieved employing a 5-fold cross validation (5-CV) procedure with stratification for fold generation. Several metrics were used to evaluate the classification models in the reported experiments. First, at the subtype level, precision (Prec), recall (Rec) and MCC were again used to evaluate the binary classifier for each subtype. At the global level, the quality of the multi-class models was evaluated using classification accuracy, and MCC was used for multi-class classifiers. SVM classification again used the LIBSVM library and an RBF kernel, whose parameters C and γ are adapted by means of a grid search.

Data preparation

Segmentation of Structural Sequence Domains Class C GPCRs have a common complex structure due to their transmembrane location: An extracellular domain comprising the N-terminus and 3 extracellular loops (EL), the 7TM, and an intracellular domain consisting of three intracellular loops (IL) and the C-terminus. Complete sequences, in accordance to this catalogue of structural domains, were partitioned into 15 segments. For this, the Phobius transmembrane detection tool [127] was used.

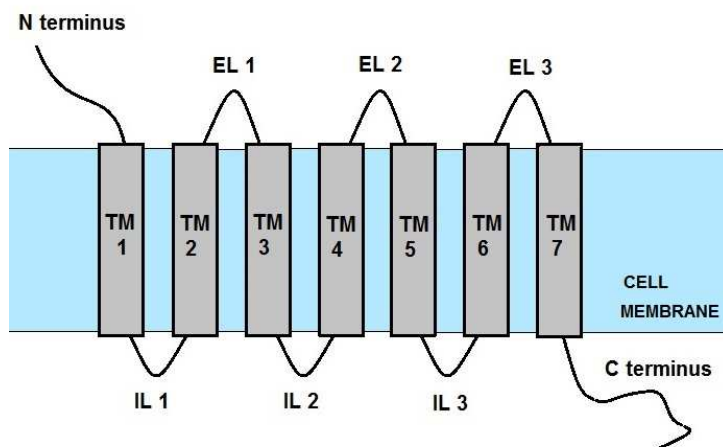


Figure 7.2.1.: Graphical representation of the common structure of GPCRs.

Alignment-free sequence transformations The use of the supervised classification models aforementioned requires transforming the unaligned AA primary sequences of varying length into fixed-size matricial representations. In the research reported in previous chapters, we used transformations based on the physicochemical properties of the AAs that have widely been used in proteomics research[49, 48]. In the current study, we use transformations that have their foundations in the field of symbolic language analysis instead. They treat protein sequences as text from a 20 AA alphabet [74, 76]. Here, short sequence fragments known as n -grams, are understood as “words”. In [50], a successful application of class A GPCR classification using text classification methods was reported. This study used a discretization of n -gram features. In our research, we followed a similar strategy and calculated the relative frequency of occurrence of n -grams of sizes one and two that we call, in turn, AA and Digram transformations. These n -gram-based transformations achieved relatively high classification performances in previous research for the analysis of the complete sequences of the original dataset [10]. Now, we go one step further and do not only calculate the frequencies of AA and Digram for all sequence segments (called *appended frequencies*), but also the *accumulated frequencies*, which are calculated as the occurrence of AA or Digram in all the segments under study divided by the sum of the lengths of these segments.

7.2.2. Analysis of the discrimination capability of segments from the 2011 dataset

Results

Segmentation of Structural Sequence Domains As previously mentioned Class C of GPCRs in the 2011 dataset is subdivided into seven subtypes: Metabotropic Glutamate (mG) receptors, Calcium sensing (CS), GABA-B (GB), Vomeronasal (VN), Pheromone (Ph), Odorant (Od) and Taste (Ta). The analyzed data set from version 11.3.4, released on March 2011, contains a total of 1,510 sequences from those seven subtypes. In this study we limited our analyses to the subset of 1,252 sequences (approximately 83% of the total) that include information of the complete 7-TM domain. The distribution of sequences per subtype both for the original data set and for the subset comprising only sequences with complete 7-TM structure are shown in Table 7.1.

Table 7.1.: Number of sequences per subtype available in the original data set and in the subset of sequences with complete 7-TM structure.

Class C subtype	# sequ. of dataset	# sequ. compl. 7-TM structure
mG	351	282
CS	48	45
GB	208	156
VN	344	293
Ph	392	323
Od	102	90
Ta	65	62
	1510	1252

Table 7.2 summarizes some general information about the lengths (in number of AAs) of the 15 segments of the sequences for the 2011 class C dataset.

Classifier performance comparison with complete sequences As stated in section 7.2.1, we first used several supervised models for the classification of the complete sequences in order to select the most adequate classifier. Table 7.3 shows the classification performance for the different classifiers (best results are highlighted in bold). The results reveal that the best classification performance was achieved by SVM, both for the AA and Digram transformations, in comparison to RF and NB. For this reason, SVM was used in the subsequent experiments.

Table 7.2.: Statistical information in reference to the length of the segments.

Segment	Mean	Min	Max	StDev
Complete Sequence	861.7	250	1,768	181
N-terminus	532.2	6	1,502	148.3
EL1	11.6	5	329	10.4
EL2	27	5	70	10.4
EL3	9	5	31	3.9
TM1	24.7	16	34	1.9
TM2	21.8	17	31	1.7
TM3	23.5	17	34	2.3
TM4	22.3	18	33	2.9
TM5	23.5	17	34	2.3
TM6	21.3	17	27	1.3
TM7	23.6	16	31	1.6
IL1	17	6	567	39.9
IL2	18.9	11	69	4.2
IL3	11.9	6	85	3.3
C-terminus	73	0	1,044	113

Table 7.4 details the underlying subtype classification results by reporting the per-subtype *Precision*, *Recall* and MCC obtained with the SVM classifier from the Digram data representation. The best results were obtained for subtypes MG, CS, GB and Ta, while the results for subtypes VN, Ph and Od were less accurate. The Od subtype, in particular, yielded very poor results. Overall, these results are, in any case, in line with those obtained in previous research ([3], [10]) and presented in Chapter 4 of this document.

Experiments with topological sequence segments The experiments reported in this section concern the SVM classification models built for the different topological segments and their combinations. Table 7.5 shows the classification results for the segments in the extracellular domain. Table 7.6 corresponds to the 7TM, and table 7.7, in turn, to the four intracellular regions IL1, IL2, IL3 and the C-terminus. Table 7.8, on the other hand, summarizes the classification results for the N-terminus combined with the 7TM region. Finally, Table 7.9 shows the classification results for all 15 segments of the complete sequence. Each table displays the name of the segments considered in the experiment, the size of the feature set and the classification performance as measured by MCC and accuracy.

Table 7.3.: Classification results for the complete sequences according to classifier.

	AA			Digram		
Classifier	N	MCC	Accu	Size	MCC	Accu
NB	20	0.625	0.703	400	0.792	0.834
RF	20	0.657	0.726	400	0.656	0.724
SVM	20	0.838	0.873	400	0.917	0.934

Table 7.4.: Subtype classification results achieved with SVM from the Digram data transformation.

Subtype	MCC	Prec	Recall
mG	0.946	0.975	0.949
CS	0.951	0.911	0.927
GB	1.0	0.981	0.989
VN	0.936	0.932	0.913
Ph	0.897	0.922	0.875
Od	0.810	0.675	0.722
Ta	1.0	1.0	1.0

Subtype specific classification results of topological sequence segments

In this section, we extend the previous information by reporting the per-subtype classification results for the sequence segments (and its combinations) found to perform best as detailed in the previous sub-section. Table 7.10 shows these subtype classification results for the concatenation of all 15 segments (MCC=0.914, Accu=0.932), the N-terminus (MCC=0.901, Accu=0.92), the extracelullar segments, i.e. N-terminus +EL (MCC=0.901, Accu=0.921), the N-terminus + 7TM (MCC=0.909, Accu =0.928), the 7TM segments (MCC=0.873, Accu=0.902) and the intracelullar segments, i.e. IL+C-terminus (MCC=0.88, Accu=0.906). For each dataset, the types of transformation and frequency are reported in the table.

Discussion

The results of our experiments for the sequence segments and their combinations neatly reveal a pattern of progressive deterioration of classification performance as we remove more parts of the sequence. It is nevertheless remarkable that the classification performance never decreases below 0.75 (neither in MCC nor in accuracy), even for very short segments, and seldom below 0.8. These results thus reveal a notable conservation of the subtype discriminability capabilities throughout the

Table 7.5.: Classification results for the extracellular segments.

Segments	AA			Digram		
	N	MCC	Accu	Size	MCC	Accu
N-terminus	20	0.792	0.835	400	0.901	0.920
EL1	20	0.802	0.842	390	0.786	0.831
EL2	20	0.798	0.839	386	0.825	0.861
EL3	20	0.779	0.825	327	0.769	0.816
All EL appended freq.	60	0.839	0.873	1103	0.873	0.880
All EL accum. freq.	20	0.804	0.845	398	0.844	0.875
(Nterm + EL) app. freq.	80	0.878	0.904	1502	0.889	0.912
(Nterm + EL) accum. freq.	20	0.8089	0.849	400	0.901	0.921

sequence.

For the entire sequence, the best classification was found for the Digram representation, which yielded an MCC of 0.917 and an accuracy of 0.934, similar performance to that of its partition into 15 segments, with an MCC of 0.914 and accuracy of 0.932 for the Digram representation and accumulated frequencies (see table 7.9). Note that by using the segmentation of the entire sequence, the classification results of the AA transformation were improved clearly as the entire sequence achieved an MCC of 0.838 and accuracy of 0.873 using 20 attributes, while the appended frequency of the 15 segments yielded an MCC of 0.905 and accuracy of 0.925 using 300 attributes. This result validates the approach consisting on the combination of complete sequence segmentation and use of appended frequencies.

The analysis of the extracellular segments revealed that the classification performance using the N-terminus alone or combined with the extracellular loops (see table 7.5) decreases just over one percentage point, both in MCC and accuracy when compared to the performance of the complete sequence and the Digram transformation. The combination of the N-terminus with the 7TM provided similar classification performances as well (see table 7.8).

The experiments corresponding to the extracellular loops, transmembrane and intracellular segments show less accurate classification compared to those of the entire sequence or the N-terminus. At large, the combination of topologically-alike segments improves the classification results obtained using single segments (with the aforementioned exception of the N-terminus). As well note that some very short sequence segments such as IL2, EL2, TM3 and TM4 (several of them comprising no more than 2.2% of the sequence) barely decrease more than 6% in classification performance when compared with the best results. This is a somewhat

Table 7.6.: Classification results for the transmembrane segments.

	AA			Digram		
Segments	Size	MCC	Accu	Size	MCC	Accu
TM1	20	0.741	0.794	321	0.778	0.823
TM2	20	0.809	0.850	298	0.806	0.847
TM3	20	0.829	0.866	290	0.846	0.878
TM4	20	0.776	0.822	320	0.822	0.860
TM5	20	0.8181	0.859	293	0.817	0.856
TM6	20	0.794	0.836	262	0.81	0.848
TM7	20	0.755	0.808	281	0.801	0.843
TM append. frequency	140	0.873	0.902	2066	0.871	0.900
TM accum. frequency	20	0.847	0.879	384	0.864	0.894

Table 7.7.: Classification results for the intracellular segments.

	AA			Digram		
Segments	N	MCC	Accu	Size	MCC	Accu
IL1	20	0.777	0.825	398	0.739	0.795
IL2	20	0.815	0.853	388	0.837	0.872
IL3	20	0.817	0.857	304	0.789	0.834
C-terminus	20	0.74	0.793	400	0.753	0.805
(IL+ C-term.) append. freq.	80	0.880	0.906	1490	0.874	0.895
(IL + C-term.) accum. freq.	20	0.795	0.837	400	0.854	0.885

Table 7.8.: Classification results for the N-terminus concatenated with the 7TM regions.

	AA			Digram		
Segments	N	MCC	Accu	N	MCC	Accu
appended frequency	160	0.897	0.919	2467	0.889	0.915
accumulated frequency	20	0.830	0.866	400	0.909	0.928

Table 7.9.: Classification results for the concatenation of all 15 segments.

	AA			Digram		
Segments	N	MCC	Accu	N	MCC	Accu
appended frequency	300	0.905	0.925	5058	0.888	0.911
accumulated frequency	20	0.840	0.875	400	0.914	0.932

surprising outcome that indicates that subtype differences are deeply embedded even in such small segments.

Regarding the type of transformation, Digram yielded, in general, the best results with two interesting exceptions, namely for the 7TM regions and the IL + C-terminus for the appended frequencies, for which the AA transformation yielded better results. The comparison between the use of appended frequencies and accumulated frequencies reveals that the former achieve better results with the AA transformation, whereas the latter perform better with Digram.

The per-subtype classification results reported in Table 7.10 are consistent with the results obtained for the entire sequence (See table 7.4), as all datasets achieve better results for subtypes MG, CS, GB and Ta, while subtypes Vn, Ph and Od perform the worst.

A detailed comparison of the subtype classification results shows that the entire sequence and the concatenation of its 15 segments provide the best performance for subtypes GB (MCC=0.989), Vn (MCC=0.913), Ph (MCC=0.875) and Ta (MCC=1.0). In turn, the best results for mG were found for the entire sequence (MCC=0.953) and N-terminus + EL (MCC=0.954). For subtype CS the best result was found for the N-terminus (MCC=0.952), while Od performed best for the combination of N-terminus + 7TM (MCC=0.764).

The overall good behavior of those sequences including the N-terminus is consistent with the fact that this domain contains the binding sites for the endogenous ligands responsible for the activation of class C GPCRs. Thus, the AAs present in the N-terminus determine the recognition of glutamate in mG receptors, GABA in GABA-B receptor, Ca^{2+} in CS receptor, *etcetera*. As a consequence, the N-terminus conveys most of the discriminatory elements for the classification of GPCR class C sequences. However, GPCRs and particularly their class C are complex entities both at the structural and functional levels. GPCRs are allosteric machines and the binding sites for the transducer G proteins are located at the intracellular part of the receptors far away from the ligand binding sites. This may explain the contribution of the ILs in our analysis. Moreover, the 7TM domain needs to be activated for G protein binding and then contributions of this

Table 7.10.: Subtype classification results for different sequence segments and transformation as described in the header. MCC best results over segment choices for each subtype shown in bold.

	Concaten. 15 segments			N-terminus		
	(Digram accum. frequ.)			(Digram)		
Subtype	Prec	Recall	MCC	Prec	Recall	MCC
mG	0.947	0.982	0.953	0.962	0.951	0.943
CS	0.951	0.933	0.939	1.0	0.911	0.952
GB	1.0	0.981	0.989	1.0	0.968	0.982
VN	0.939	0.929	0.913	0.919	0.918	0.893
Ph	0.894	0.922	0.875	0.88	0.916	0.859
Od	0.853	0.688	0.722	0.751	0.725	0.712
Ta	1.0	1.0	1.0	1.0	0.967	0.982
	N-terminus + EL			N-terminus + 7TM		
	(Digram accum. frequ.)			(Digram app. frequ.)		
Subtype	Prec	Recall	MCC	Prec	Recall	MCC
mG	0.968	0.961	0.954	0.922	0.986	0.939
CS	0.980	0.889	0.928	0.933	0.889	0.906
GB	0.993	0.974	0.982	1.0	0.962	0.978
VN	0.912	0.908	0.882	0.917	0.894	0.877
Ph	0.865	0.919	0.851	0.878	0.904	0.851
Od	0.752	0.688	0.70	0.873	0.70	0.764
Ta	1.0	0.9372	0.966	1.0	0.983	0.991
	Transmembrane			IL + C-terminus		
	(AA app. frequ.)			(AA app. frequ.)		
Subtype	Prec	Recall	MCC	Prec	Recall	MCC
mG	0.926	0.986	0.94	0.953	0.975	0.953
CS	0.899	0.933	0.912	0.939	0.889	0.908
GB	1.0	0.968	0.982	0.982	0.9811	0.978
VN	0.89	0.894	0.859	0.879	0.918	0.867
Ph	0.873	0.883	0.833	0.884	0.898	0.85
Od	0.667	0.5	0.549	0.75	0.513	0.592
Ta	1.0	0.954	0.974	0.986	0.985	0.984

structural domain for sequence classification are expected. Inasmuch as allosteric cooperativity interactions between the 7TM and VFT domains have been also reported [128], it is expected that segments including these domains appear in our study. Finally, ELs are involved in 7TM domain flexibility and cooperativity interactions, which justify their putative discriminative power.

As a whole, these results provide a complete and detailed landscape of the relative capabilities of different sequence segments (from different GPCR domains and in different combinations) in the task of discriminating between the seven subtypes of class C GPCRs. This detailed landscape should help database biocurators in their tasks.

Conclusions

The research reported in this section is based on the web-accessible and public protein databases of the GPCRdb consortium. Biocurators of this type of databases face the non-trivial challenge of unambiguously identifying and characterizing GPCRs. In this database, receptors are characterized according to subtype labels at different levels of organization. In previous research [12], the analysis of the N-terminus of the extracellular domain provided some preliminary evidence of the potential use of individual domains of complete class C GPCR sequences as the foundation for subtype classification.

In this research, we have performed a systematic analysis of the classification performance of each of the individual sequence segments in which the sequence can be divided in each of its structural domains, as well as the performance of several of their combinations. The experimental results revealed that none of them reached the classification performance of the complete sequence or the concatenation of its 15 constituent segments. However the segments of the extracellular domain, the N-terminus in combination with the 7TM and, to some degree, the intracellular domain have all performed almost as well as the complete sequence. The identification of these most discriminative segments should be the starting point for future work focusing on these separate regions. Such future research should involve feature selection starting from these segments as a way to discover specific motifs with subtype discriminative capabilities and potential functional roles.

7.2.3. Analysis of the discrimination capability of segments from the September 2016 dataset

Results

Segmentation of Structural Sequence Domains The September 2016 version is subdivided in only four subtypes: Metabotropic Glutamate (mG) receptors, Calcium sensing (CS), GABA-B (GB) and Taste (Ta). The analyzed data set contains a total of 936 sequences from those four subtypes. We limited our analyses to the subset of 922 sequences (approximately 98% of the total) that include information of the complete 7-TM domain. The distribution of sequences per subtype, both for the original dataset and for the subset comprising only sequences with complete 7-TM structure, are shown in Table 7.11.

Table 7.11.: Number of sequences per subtype available in the original data set and in the subset of sequences with complete 7-TM structure.

Class C subtype	# sequ. of dataset	# sequ. compl. 7-TM structure
mG	516	505
CS	103	103
GB	89	88
Ta	228	226
	936	922

Complete sequences, in accordance to this catalogue of structural domains, were again partitioned into 15 segments using the Phobius transmembrane detection tool. Table 7.12 summarizes some general information about the lengths (in number of AAs) of these segments for the September 2016 class C dataset.

Classifier performance comparison with complete sequences As stated in Section 7.2.1, we first used several supervised models for the classification of the complete sequences in order to select the most adequate classifier. Table 7.13 shows the classification performance for the different classifiers (best results are highlighted in bold). The results reveal that all three classifiers perform quite good, but nevertheless the best classification performance was achieved by SVM, both for the AA and Digram transformations, in comparison to RF and NB. For this reason, SVM was used in the subsequent experiments.

Table 7.14 details the underlying subtype classification results by reporting the per-subtype *Precision*, *Recall* and MCC obtained with the SVM classifier from

Table 7.12.: Statistical information in reference to the length of the segments.

Segment	Mean	Min	Max	StDev
Complete Sequence	927.88	453	1992	179.19
N-terminus	545.56	133	1070	114.44
EL1	11.69	5	593	19.55
EL2	26.85	5	70	4.95
EL3	7.86	5	25	5.53
TM1	24.67	18	28	1.75
TM2	21.72	18	31	1.6
TM3	21.86	17	27	2.38
TM4	23.78	19	29	2.76
TM5	22.16	18	31	2.06
TM6	22.35	17	28	1.76
TM7	22.92	18	28	2.1
IL1	13.7	6	284	21,37
IL2	19.8	11	56	3,44
IL3	11.94	11	69	3.59
C-terminus	130.48	0	610	140.21

the Digram data representation. For all four subtypes very accurate classification results are achieved.

Experiments with topological sequence segments The experiments reported in this section concern the SVM classification models built for the different topological segments and their combinations. Table 7.15 shows the classification results for the segments in the extracellular domain. Table 7.16 correspond to the 7TM, and table 7.17, in turn, to the four intracellular regions IL1, IL2, IL3 and the C-terminus. Table 7.18, on the other hand, summarizes the classification results for the N-terminus combined with the 7TM region. Finally, table 7.19 shows the classification results for all 15 segments of the complete sequence. Each table displays the name of the segments considered in the experiment, the size of the feature set and the classification performance as measured by MCC and accuracy.

Table 7.13.: Classification results for the complete sequences according to classifier.

	AA			Digram		
Classifier	N	MCC	Accu	Size	MCC	Accu
NB	20	0.933	0.959	400	0.983	0.989
RF	20	0.945	0.967	400	0.923	0.948
SVM	20	0.980	0.988	400	0.994	0.996

Table 7.14.: Subtype classification results achieved with SVM from the Digram data transformation.

Subtype	MCC	Prec	Recall
mG	0.996	0.998	0.998
CS	0.984	0.981	0.991
GB	0.994	1.0	0.989
Ta	0.994	0.996	0.996

Discussion

The experimental results show very accurate classification results for all segments. Only a minor deterioration of classification happens on account of the elimination of sequence segments. Nevertheless it is remarkable that the performance never drops below 0.941 (measured as MCC), even for very small segments, such as the intracellular loops or transmembrane regions, what implies a noticeable conservation of the discrimination capability for all sequence segments.

The best classification in our experiments using the entire sequences was found for the Digram representation with an accuracy of 0.996 and MCC of 0.994 (see table 7.13). The combination of the intracellular loops and the C-terminus even improves this result slightly with an accuracy of 0.997 and MCC of 0.995. Also the combination of the 7TM regions or the combination of 7TM segments and the N-terminus achieve nearly the same result with an accuracy of 0.996 and MCC of 0.993 (see table 7.16 and 7.18). As well the concatenation of all 15 segments yield very close classification results with an accuracy of 0.996 and MCC of 0.994 (see table 7.19).

The N-terminus by itself, or in combination with the extracellular loops (see Table 7.15), are less accurate than those of the entire sequence for both the AA and

Table 7.15.: Classification results for the extracellular segments.

Segments	AA			Digram		
	N	MCC	Accu	Size	MCC	Accu
N-terminus	20	0.98	0.988	400	0.986	0.991
EL1	20	0.941	0.963	363	0.956	0.971
EL2	20	0.961	0.975	373	0.964	0.975
EL3	19	0.966	0.978	223	0.962	0.976
All EL appended freq.	59	0.976	0.985	959	0.973	0.983
All EL accum. freq.	20	0.974	0.983	395	0.972	0.982
(Nterm + EL) app. freq.	79	0.986	0.991	1359	0.984	0.989
(Nterm + EL) accum. freq.	20	0.968	0.980	400	0.976	0.984

Digram transformation as they yield an accuracy of 0.991 and MCC of 0.986.

Typically, the combination of topologically-alike segments improves the classification results of single segments. It is noteworthy that even very small segments such as EL1, TM1 and TM4 (some of them including on average no more than 2.2% of the AAs of the sequence) just decrease at most 5% in classification performance as compared with the best results.

Conclusions

Preliminary research hinted the potential use of separated domains of complete class C GPCR sequences as the basis for subtype classification. In this section, we have reported the results of a systematic analysis of the performance of each of the individual sequence segments and some of their combinations for the 2016 version dataset. The results are consistent with those reported for the former 2011 version dataset (see section 7.2.2) as the best classification results were found for the complete sequence. Nevertheless the extracellular domain, the combination of the N-terminus and 7TM and, to some extent, the intracellular domain have all performed almost as well as the entire sequence. This, by itself, allows us to focus our work on the most discriminative segments.

Table 7.16.: Classification results for the transmembrane segments.

Segments	AA			Digram		
	Size	MCC	Accu	Size	MCC	Accu
TM1	20	0.947	0.965	236	0.962	0.975
TM2	20	0.969	0.979	216	0.986	0.99
TM3	20	0.993	0.995	237	0.987	0.991
TM4	20	0.972	0.982	254	0.954	0.972
TM5	20	0.985	0.991	221	0.985	0.99
TM6	20	0.978	0.985	185	0.985	0.99
TM7	20	0.988	0.992	222	0.987	0.992
TM append. frequency	140	0.993	0.996	1571	0.991	0.993
TM accum. frequency	20	0.979	0.987	344	0.991	0.993

Table 7.17.: Classification results for the intracellular segments.

Segments	AA			Digram		
	N	MCC	Accu	Size	MCC	Accu
IL1	20	0.964	0.977	392	0.962	0.975
IL2	20	0.981	0.987	334	0.982	0.988
IL3	20	0.987	0.991	218	0.983	0.988
C-terminus	20	0.949	0.969	398	0.981	0.988
(IL+ C-term.) append. freq.	80	0.995	0.997	1342	0.985	0.989
(IL + C-term.) accum. freq.	20	0.984	0.99	400	0.988	0.992

Table 7.18.: Classification results for the N-terminus concatenated with the 7TM regions.

Segments	AA			Digram		
	N	MCC	Accu	N	MCC	Accu
appended frequency	160	0.993	0.996	1971	0.992	0.995
accumulated frequency	20	0.986	0.991	400	0.992	0.995

Table 7.19.: Classification results for the concatenation of all 15 segments.

Segments	AA			Digram		
	N	MCC	Accu	N	MCC	Accu
appended frequency	299	0.994	0.996	4272	0.991	0.993
accumulated frequency	20	0.986	0.991	400	0.992	0.995

8. Discovering class C GPCR motifs

8.1. Introduction

In this chapter, we build on the research reported in previous ones and describe the use of FS techniques to build SVM-based classification models from selected receptor subsequences described as n -grams. We show that this approach to classification is useful for identifying class C GPCR subtype specific motifs. The n -gram transformation of the GPCR sequences is likely to yield many features that are not relevant in terms of class C subtype discrimination. These irrelevant n -gram frequencies may have a negative impact (or at best a negligible one) in this classification process and, therefore, we aim to investigate whether a subset of relevant frequencies retains the subtype classification capabilities. Indirectly, we also want to investigate the selected n -grams for hitherto unknown signature motifs. One criterion of significance is their statistical or informative performance, which is related to biological significance [129]. It has been suggested that motif over-representation maybe due to evolutionary preservation of sequence segments, signalling their structural and functional roles [15]. This should make n -gram frequencies informative measures in terms of functionality exploration.

We present several experiments in which different dimensionality reduction methods are applied to the class C GPCR dataset. Several FS approaches are used: sequential forward feature selection (SFFS) with an SVM-classifier and filter methods using univariant metrics such as t-test or χ^2 measures. Moreover, we conduct experiments using different classification models: the one-*vs*-one approach to filter out subtype discriminating pattern, i.e. those patterns which best distinguish between subtypes, and the one-*vs*-all approach to detect subtype specific pattern, i.e. those patterns which are most characteristic for a subtype. Our experiments also consider the complete receptor sequence or specific receptor segments, such as the N-terminus domain, for example.

8.2. Feature selection used for the identification of subtype-discriminating n-grams

8.2.1. Experiments on the complete sequences

Experimental settings

In this experiment we analyze the complete receptor sequences from the 2011 version class C GPCR dataset. We use different FS techniques to build SVM-based classification models from n -gram frequency presentations in order to find subtype discriminating n -grams. First, we built classification models with n -grams for each of the three alphabets (AA, SEZ, DAV) for n -grams with lengths up to three.

To improve the quality and interpretability of the classification models, different dimensionality reduction methods that discard irrelevant and redundant features and retain a subset of highly discriminative features are used. In this experiments two FS approaches are applied, namely a filter method computing two-sample t-tests among the seven class C GPCR subtypes using a one-*vs*-one approach and a SFFS approach with a SVM-classifier:

- A two-sample t-test was used to evaluate the discriminating power of each feature as a filtering approach. This univariate statistical test analyzes whether there are foundations to consider two independent samples as coming from populations (normal distributions) with unequal means by analyzing the values of the given feature. In our case, we used t-tests with 0.01 confidence. If the t-test suggested that this hypothesis was true (i.e. the null hypothesis was rejected), the feature was considered to significantly distinguish between the two different subtypes of class C GPCRs. As we face a multi-class classification problem, the t-test results were examined for the 21 feasible two-class combinations of the 7 class C subtypes. We decided to calculate the two-sample t-test values at this detail because the multi-class LIBSVM implementation internally performs a comparison of the data between each class (one-*vs*-one implementation). Therefore, the t-test exactly evaluates the data considered in each binary classifier, making the ranking of the features possible according to their overall significance (i.e., in how many binary classifiers a feature is significant). A more robust estimate of the proportion of significant features by means of the q -value approach [92] is not applied, as the t-test selection constitutes a coarse feature filter as first step of the FS approach.

- A SFFS algorithm was used to find the reduced set of features that best discriminated the data subtypes. This kind of algorithm is a so called wrapper method, where the classification model search is performed within the subset feature search. This algorithm starts from an empty candidate feature set and adds, in each iteration, the feature which most improves the accuracy (i.e., that which minimizes the misclassification rate). The algorithm uses a SVM classifier in which the accuracy is evaluated using a 5-CV to test the candidate feature set. The algorithm stops when the addition of a further feature does not increase the accuracy over a threshold set at $1e^{-6}$.
- A combination of a t-test filter in a first step followed subsequently by an SFFS approach.

The classification performance of the results is measured by means of the accuracy and MCC for multiclass classification. The subtype results are evaluated by means of the MCC, precision and recall for binary classification. All experiments are conducted using 5-CV with stratified folds.

Results

From the comparison of different supervised classifiers for the analysis of the n -gram transformed datasets (See Table 4.4 reported in section 4.1.3), it has been shown that SVM outperforms the rest of classifiers and, therefore, is the most adequate choice to use with the n -gram frequency representations in subsequent experiments.

Table 8.1 again summarizes the classification results obtained only with SVM classifiers as well as the feature size of the different n -gram representations. Note that each element in each alphabet is itself considered as a 1-gram, regardless the number of constituent AAs. Obviously, the size of the n -gram feature set increases significantly with the size of the alphabet. Results are shown for 1-grams, 2-grams, and the combination of both (1,2-gram).

Table 8.1.: N -gram classification results for the different alphabets without FS, where N is the size of a feature set and Accu stands for classification accuracy (ratio of correctly classified sequences to all sequences).

N-gram	AA			SEZ			DAV		
	N	Accu	MCC	N	Accu	MCC	N	Accu	MCC
1-gram	20	0.88	0.84	11	0.815	0.763	9	0.791	0.736
2-gram	400	0.932	0.914	121	0.923	0.903	81	0.911	0.888
1,2-gram	420	0.934	0.917	132	0.925	0.906	90	0.919	0.898

The combination of 1 and 2-grams of the AA alphabet reached the best classification results with an accuracy of 0.934 and MCC of 0.917. The construction of an SVM model from 3-grams for all three alphabets was unsuccessful, probably due to the existence of a large set of irrelevant features. This was the primary reason behind the decision of applying FS as part of the classification process. The dimensionality reduction was implemented in this experiment using two different FS approaches: SFFS with an SVM-classifier and a filter method computing two-sample t-tests among the class C GPCR subtypes.

Sequential Forward Feature Selection Table 8.2 shows the classification results when SFFS was performed on each n -gram dataset. For each alphabet (AA, SEZ and DAV), this table shows a comparison between the original size of the n -grams (N) and the number of selected features found by the algorithm, as well as the corresponding classification accuracy.

The experiments show that the FS algorithm was successful, with only one exception: in the case of the 1,2,3-gram feature set (combination of all n -grams) of the AA-alphabet: due to the large number of features, the computational cost of the SFFS algorithm is too high. In fact, this was the result that prompted us to investigate a classifier-independent filter FS method that could provide us with a first rough selection of features to be used as a preliminary step to a subsequent process of forward FS.

Table 8.2.: N -gram classification results using SFFS, for the three different alphabets

N-gram	AA			SEZ			DAV		
	N	FS	Accu	N	FS	Accu	N	FS	Accu
1-gram	20	17	0.88	11	-	-	9	-	-
2-gram	400	48	0.93	121	25	0.906	81	31	0.9
1,2-gram	420	54	0.926	132	37	0.916	90	42	0.92
1,2,3-gram	8420	-	-	1131	34	0.925	818	34	0.923

t-Test Filtering In order to handle the 1,2,3-gram feature sets, which, due to their size, were either impossible or very difficult to use in the previous methods, we decided to use the t-test filtering method to establish a ranking of the features. Table 8.3 shows this ranking according to the overall significance of the attributes. This means that, for each alphabet, we counted how many features were significant (column N) in at least 20,19,18, etc. two-class tests. The Accu values shown for each subset are the classification accuracies of a SVM-classifier built on each feature set.

These results provide evidence of the usefulness of this simple ranking, as we were able to find subsets that outperform the classification accuracies obtained with the previous methods. For example, the 1,2,3-gram representation of the AA alphabet achieves an accuracy of 0.943 with 585 attributes, whereas the 2-gram representation achieves a 0.93. In the case of the SEZ alphabet, an accuracy of 0.943 was obtained with this filtered 1,2,3-gram representation, as compared to 0.926 with the 2-gram representation. Using the DAV alphabet, we found a subset with 238 features that yielded a 0.933 accuracy, whereas the 1,2,3-gram representation with SFFS yielded a 0.92.

Table 8.3.: Classification results with t-test-based subset selection, with subsets of features that are significant in a given number of t-tests, from 20 down to 12

	AA		SEZ		DAV	
SIGNIF	N	Accu	N	Accu	N	Accu
20	1	0.37	2	0.5	0	-
19	15	0.88	8	0.77	10	0.83
18	49	0.931	39	0.9	23	0.88
17	105	0.933	79	0.922	58	0.91
16	212	0.937	149	0.93	99	0.92
15	357	0.936	253	0.936	164	0.926
14	585	0.943	386	0.935	238	0.933
13	909	0.937	505	0.943	325	0.93
12	1284	0.942	633	0.94	429	0.927

t-Test Filtering and Sequential Forward Feature Selection The filtering method described in the previous section found feature subsets with high classification accuracy. Nevertheless, given their high dimensionality, we decided to apply the SFFS algorithm to these subsets as a subsequent dimensionality reduction step. Table 8.4 shows the results of applying SFFS starting from the n -gram subset reported in the last row of Table 8.3 (features relevant in at least 12 classifiers), for each alphabet. The initial number of features (FEAT), the number of selected features (N) and the corresponding classification accuracies are shown. Forward selection was quite successful at reducing the number of attributes while retaining an accuracy of approximately 0.94 in all three cases.

Table 8.4.: Classification results with FS on top of t-test-based selection for a subset solution in which features are significant in at least 12 of the 21 t-tests.

AA			SEZ			DAV		
FEAT	N	Accu	FEAT	N	Accu	FEAT	N	Accu
1284	49	0.939	633	59	0.939	429	60	0.94

Discussion

Classification from n -grams with and without feature selection The results reported in Table 8.3 provide clear evidence of the usefulness of the t-test-based simple feature ranking method, as parsimonious feature subsets that outperform the classification accuracies obtained without FS or with forward selection on its own were found. For example, the 1,2,3-gram representation of the AA alphabet achieves an accuracy of 0.943 with 585 attributes, improving on the 0.930 accuracy obtained directly with the 2-gram representation using only forward selection (as reported in Table 8.2). In the case of the SEZ alphabet, the same 0.943 accuracy was obtained with this filtered 1,2,3-gram representation with 505 features; this has again to be compared to the 0.926 obtained with the 2-gram representation (Table 8.1) and the 0.925 obtained with the 1,2,3-gram representation (Table 8.2). Using the DAV alphabet, we found a subset with 238 features that yielded a 0.933 accuracy, whereas the 1,2,3-gram representation with forward selection yielded a 0.920 (Table 8.2).

Nevertheless, the filter selection method on its own still renders rather high-dimensional optimal solutions and the slight classification improvement it generates might not be enough to counter-balance the complexity of the solution. In fact, the most interesting results, as reported in Table 8.4, come from the application of the classifier-dependent forward selection to the results of the filter method. Results show that this approach was quite successful at reducing the number of attributes by as much as 96% while retaining an accuracy in the area of 0.94 for all three alphabets.

Overall, the experimental results reported in the previous section support the interest of using FS on the analyzed n -gram data: data dimensionality has been notably reduced without compromising classification quality. Forward selection has been shown to be an effective method, although it is computationally too costly when the size of the feature set is too high from the onset. In this situation, a fast univariate t-test-based filtering method becomes an appropriate solution to reduce the feature candidate set as a preprocessing step prior to the forward selection algorithm.

To the best of the authors' knowledge, the reported classification results are the best to date using the class C GPCRDB database, comparing favourably to those in [3, 4, 71]. These results correspond to the reduced SEZ dataset with 505 attributes (See Table 8.3) and a SVM model with parameter $C=2$, $\gamma=2^{-10}$ achieving a mean accuracy of 0.943 and a mean MCC of 0.93. Table 8.5 shows the corresponding subtype classification results, i.e. the precision, recall and MCC of each binary classification.

Table 8.5.: Subtype classification result obtained with reduced SEZ dataset with 505 attributes

Subtype	MCC	Precision	Recall
mG	0.962	0.956	0.986
CS	0.924	0.978	0.88
GB	0.997	1.0	0.995
VN	0.910	0.923	0.939
Ph	0.902	0.931	0.924
Od	0.808	0.865	0.782
Ta	0.983	1.000	1.0

Qualifying feature selection from t-test values An analysis of the t-test values (hypothesis value and p -value) allows measuring to what degree an individual feature discriminates between two classes. Test values are first analyzed to detect the 3-grams with the best discrimination capabilities. We subsequently analyze if these 3-grams may be part of larger n -grams which may also be discriminative.

The close scrutiny of the test values of the reduced feature set of the AA alphabet (See table 8.4: 49 features, including 33 3-grams, 13 2-grams and 3 1-grams) revealed that the 3-grams CSL, ITF and FSM are the most significantly discriminative.

CSL, in particular, is the most significant one according to the t-test values of 20 two-sample tests. This feature was only found not to be significant for the mG *vs.* Ph discrimination.

The ITF n -gram is deemed to be significant in 18 tests and an analysis of longer n -grams (results not reported) showed that the the ITFS 4-gram is specially discriminating, with a significant impact on the discrimination of 19 binary classifiers (i.e., all but mG *vs.* Ta and CS *vs.* Ta). Furthermore, the ITFSM 5-gram is still highly discriminative, showing significant values for 17 tests.

Another relevant 3-gram is FSM, which is significant for 18 two-class tests. An analysis of longer n -grams showed that the FSML 4-gram is highly discriminative (in 18 tests: all but mG *vs.* GB, mG *vs.* Ta and GB *vs.* Ta). The FSMLI 5-gram was also found to be significant for 15 tests.

Figure 8.2.1 shows the mean values of n -gram features CSL, ITFS, and FSML for the 7 class C GPCR subtypes.

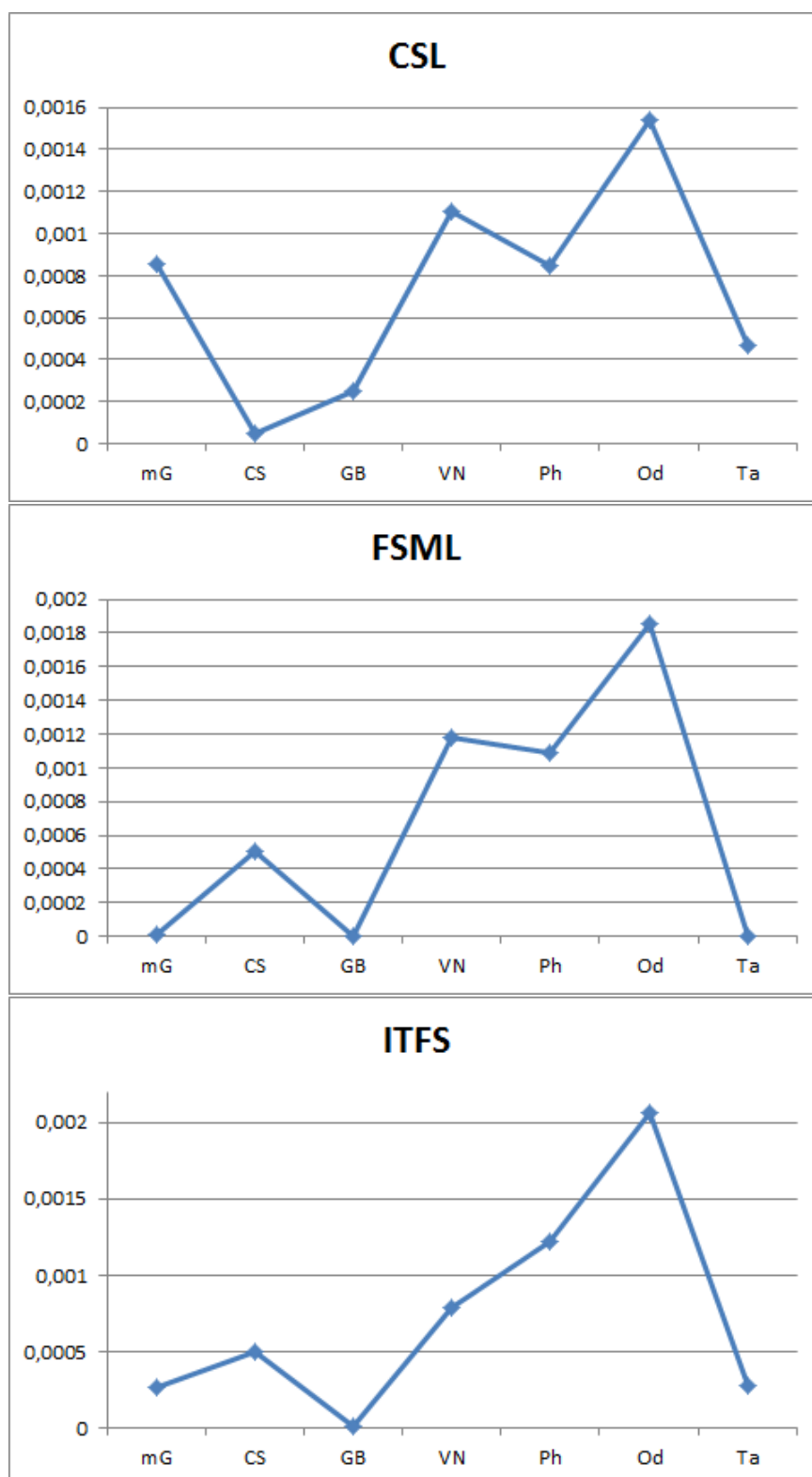


Figure 8.2.1.: Mean values of CSL (top), FSML (center) and ITFS (bottom) *N*-gram features for the 7 class C GPCR subfamilies.

Beyond mean values, a statistical analysis of the most discriminative 3-grams, CLS, ITF and FSM, revealed the existence of extreme values in some subfamily distributions, which would require a deeper analysis: Figures 8.2.2, 8.2.3 and 8.2.4 display box plots of the corresponding n -grams. The box describes the range of values between the first and the third quartiles (Q1 and Q3) with the median (Q2) as the horizontal line inside the box. The crosses are data considered to be outliers, which, in this case, are points which fall below $Q1-1.5(IQR)$ or above $Q3+1.5(IQR)$, where IQR is the interquartile range described by the box. The interval in which the data are considered not to be outliers is represented in the plot by the dashed lines stemming from the box.

The n -gram CSL (Figure 8.2.2), which was found to be discriminant in 20 two-class tests, has its maximum values for classes Od, VN, Ph and mG, whereas this n -gram is mostly non-existent in classes CS and GB. The statistical analysis of the distribution of this n -gram confirms that CSL is suitable for the description of nearly all subfamilies (except GB) as only a relative small number of outlier values exist for all of them. In subfamily GB, this n -gram is mostly non-existent, but 17% of the sequences of this subclass appear as outliers (corresponding to sequences containing this n -gram). In consequence, subfamily GB is not well represented by n -gram CSL as its distribution is not uniform. A superficial analysis of the location of the n -grams in the sequence shows that in class Od, this n -gram appears near the middle of the sequences as well as near to their end. In the case of Ta, it appears often near the beginning, while in VN it appears in all positions (beginning, middle and end).

The n -gram ITF (Figure 8.2.3) was found to be discriminant in 18 tests and has maximum values for the subfamilies Od, Ph, VN and CS. The data of the corresponding box plot confirms that this n -gram is suitable for the discrimination of these subfamilies as the existence of extreme values is quite low in these cases. For GB and Ta, this n -gram is mostly non-existent as both the median and the IQR are zero and a low number of sequences have a positive frequency of this n -gram. Despite the fact that mG also has a median and IQR with value zero, mG has to be considered a special case as its distribution has approx. 10% of outliers, which correspond to sequences containing this n -gram. Regarding the subsequence specific location of the ITF n -gram, it appears in any position (beginning, middle and end) in class Od, while in Ph and VN, it is predominantly located at the end, and in CS it is found near the middle section.

Finally, n -gram FSM (Figure 8.2.4), which was deemed significant in 18 tests, shows maximum values for subfamilies Od, VN, Ph and CS and is mostly non-existent in subfamilies mG, GB and Ta. Nevertheless, the box plot representation suggests that this n -gram describes properly Od and CS as subclasses with presence of this n -gram and GB and Ta as subfamilies not containing this n -gram.

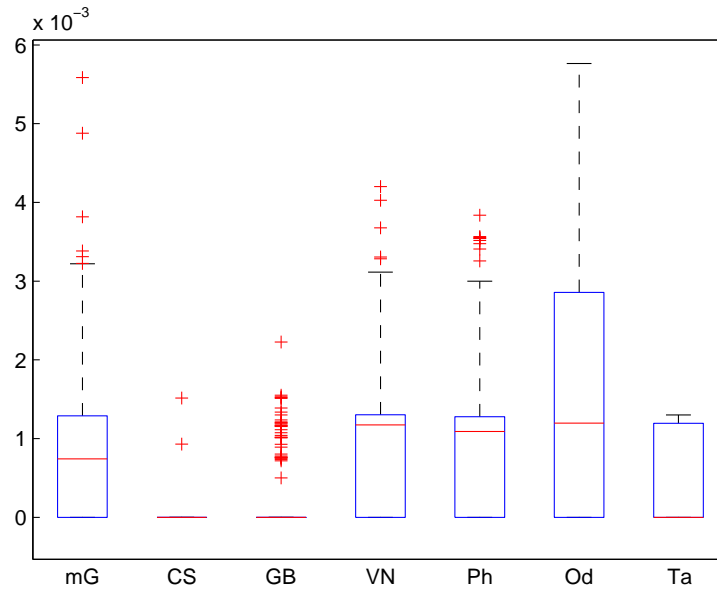


Figure 8.2.2.: Box plot of the CSL n -gram

Subfamilies mG, VN and Ph show a higher number of outliers, namely 5% (mG), 14% (VN) and 13% (Ph), which indicates that the appearance of FSM in these subfamilies is not uniform. Regarding the location of the n -gram, FSM appears in the class Od at the middle and at the end of the sequence. In the case of CS, it appears at the middle; in Vn, it appears at the end, and in Ph, both at the end and beginning.

Overall, these n -grams might be the basis for an ulterior investigation of specific motifs in class C GPCR sequences that might provide clues about ligand binding processes.

Conclusions

This experiments have addressed the problem of class C GPCR subtype discrimination according to a novel methodology that transforms the sequences according to the frequency of occurrence of the low level n -grams of different AA alphabets.

These sequence transformations generate high-dimensional data sets that are likely to include plenty of irrelevant information. For this reason, dimensionality reduction through combination of a two-sample t-test and forward FS was implemented as part of classification with SVMs.

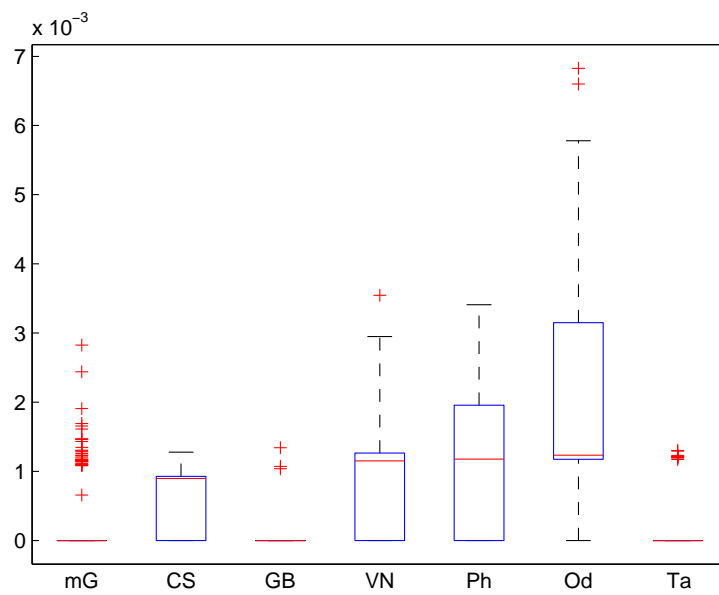


Figure 8.2.3.: Box plot of the ITF n -gram

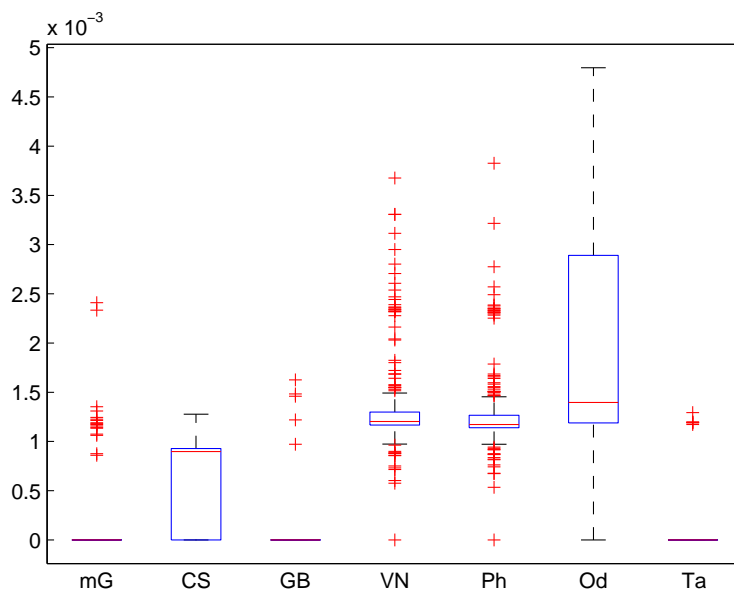


Figure 8.2.4.: Box plot of the FSM n -gram

Reduced sets of n -grams that yielded similar classification accuracies were found for each of the three transformation alphabets. These results are the best reported to date using the class C data from the 2011 version of the GPCRDB database.

The analysis of the features of the AA alphabet using the values obtained in the t-tests has provided insight about the n -grams that are best at discriminating between the GPCR subfamilies. This might be considered as preliminary evidence of the existence of subfamily-specific motifs that might reveal information about ligand binding processes. For this reason, the proposed method will be extended in future work to the analysis of larger n -grams. From this analysis, we expect to find larger n -grams that might actually be considered as potentially true subtype-specific motifs.

The analysis of the statistical distributions of the attribute values provided further insight about the nature of the analyzed data. Although the highly discriminative n -grams contributed to achieve high classification accuracy, the detected n -grams were not equally suitable to explain the data of all subfamilies. The n -grams were only appropriate to describe the distribution of the values of given subsets of subfamilies. This may be the result of the heterogeneity of some of these subfamilies. As explained in section 2.4.1, some subfamilies group nodes which are descendants from evolutionary unrelated proteins leading to separate groups. On the other hand, the data also contains overlapping data as some subclasses contain sequences which are descendants from a common ancestor. This might come to explain why, in this multi-class classification problem, the feature selection algorithm required to reach a certain number of attributes (10-30) to achieve high classification accuracies. In future work, we will address this issue by taking into account the possible subdivisions of the analyzed subfamilies.

The study of the location of n -grams in the sequence revealed that they do not appear at the same locations in different subfamilies. This discovery encourages us to apply the proposed feature selection method to separated sequence segments in order to compare n -grams specific to their subsequence specific location.

8.2.2. Experiments on the N-terminus

Experimental settings

In this experiment we analyze the N-terminus segment of the sequences of the 2011 version dataset. We base the settings of the experiment on the findings of the previous experiments: First, we focus on the N-terminus segment, which resulted nearly as discriminative as the complete sequence according to the systematic analysis of the discrimination power of the different segments of class C GPCRs

[8] (See chapter 7). Then, we use a two step FS approach, where a t-test filter and a SFFS approach are combined, which was tested in [11] and yielded the best results for the analysis on the complete sequences.

Regarding the data transformations of the primary sequences, we use very similar n -gram representations as in the previous experiment: The n -gram frequency representations are built from different alphabets, the AA alphabet and SEZ alphabet. The latter, in previous research [11], yielded a reduced dataset with very accurate classification results.

A difference in this experiment with respect to previous research is that the concept of n -gram is extended: The n -grams are, in general, contiguous specific AA subsequences of length n . For the experiments reported in the following sections, we consider a combination of *contiguous* and *rigid gap* motifs or n -grams of lengths three up to five AAs. The latter are *rigid* in the sense that there is a fixed number of gaps in between the n -gram AAs [15].

The experiments on the N-terminus comprise the following steps:

1. Data preprocessing: The primary AA sequences of the N-terminus are transformed both for the AA and SEZ alphabet to the AA and Digram frequencies, but also contiguous and rigid gap n -gram frequencies of lengths three to five are derived.
2. Comparative classification of the N-terminal domain: The classification performance by SVMs are compared between the N-terminus and the complete sequence using the AA and Digram composition for the AA and SEZ alphabet.
3. t-Test Filtering and SFFS: Given a feature space built by contiguous and rigid gap n -grams (of length three to five) the two-step dimensionality reduction approach is applied with the same experimental settings as those of the previous experiment on the complete sequence (See section 8.2.1).

Experimental results

Comparative classification of the N-terminal domain We built our SVM-based classification models using the n -grams from the N-terminus for each of the two alphabets under consideration: the complete AA and SEZ alphabet. In previous research [11], we analyzed the AA frequencies (1-grams) and digrams from the complete sequence. For comparative purposes, Table 8.6 shows the classification results, as measured by accuracy, for each alphabet using n -grams of length 1 and 2, for both approaches.

Table 8.6.: N-gram comparative classification results for the N-terminus and the complete sequence, where N is the size of a feature set and ACC stands for classification accuracy (ratio of correctly classified sequences).

	N-terminus				Complete sequence			
	AA		SEZ		AA		SEZ	
N-GRAM	N	Accu	N	Accu	N	Accu	N	Accu
1-gram	20	0.84	11	0.78	20	0.87	11	0.82
2-gram	400	0.92	121	0.91	400	0.93	121	0.93

t-Test Filtering and Sequential Forward Feature Selection Supported by the previous results, we then proceeded to apply the proposed two-step feature FS with SVM-based classification to the combination of *contiguous* and *rigid gap motifs* (n -grams) of lengths three to five. Their very high dimensionality makes them difficult to use with SVMs and, therefore, t-test filtering was used to generate a first crude ranking of features.

Table 8.7 shows this ranking according to the overall significance of the attributes. This means that, for each alphabet, we counted how many features were significant (column N) in at least 20,19,18,17, etc., subtype-vs-subtype tests (bear in mind that there are 21 possible combinations of the 7 class C subtypes). The Accu values shown for each subset are the classification accuracies of the SVM built from each feature set. For comparison, we also show, besides the results obtained with *rigid gap motifs* for the N-terminus, the corresponding results from previous research [11, 10] in which *continuous* (n -grams) of lengths one to three were calculated from the complete sequence.

The filtering method used in this experiment found feature subsets with high classification accuracy. Nevertheless, their dimensionality is still quite high, which is the reason we applied the more nuanced second step of dimensionality reduction consisting on SVM-based SFFS. Table 8.8 shows, for each alphabet, the results of applying this method starting from the n -gram subset that is significant in 16 subtype-vs-subtype problems, as reported in Table 8.7. The initial number of features (FEAT), the final number of selected features (N) and the corresponding classification accuracies are displayed.

Discussion

From Table 8.6, it seems clear that the classification analysis using only the N-Terminus almost completely retains the accuracies obtained using the complete

Table 8.7.: N-gram comparative classification results after t-test, where N is the size of a feature set and ACC stands for classification accuracy (ratio of correctly classified sequences).

	N-terminus				Complete Sequence			
	AA		SEZ		AA		SEZ	
SIGNIF	N	Accu	N	Accu	N	Accu	N	Accu
20	-	-	-	-	1	0.37	2	0.5
19	4	0.55	11	0.76	15	0.88	8	0.77
18	25	0.87	42	0.88	49	0.93	39	0.9
17	97	0.92	133	0.915	105	0.93	79	0.92
16	268	0.92	331	0.92	212	0.94	149	0.93
15	600	0.93	649	0.92	357	0.94	253	0.94
14	1187	0.93	1185	0.92	585	0.94	386	0.93

Table 8.8.: Classification results for the AA and SEZ alphabets, using SFFS starting from the first stage, t-test-based selection that is significant in 16 subtype-*vs*-subtype t-tests.

AA			SEZ		
FEAT	N	ACC	FEAT	N	ACC
268	45	0.91	331	43	0.90

sequences, specially for the digram representation. This is consistent with the fact that the VFT, included in the N-terminus, contains the orthosteric binding site that, because it differentiates between different endogenous ligands, should also help to differentiate between the different class C subtypes. From a practical viewpoint, this result potentially simplifies the search for signature motifs by restricting it to the extracellular domain, while making the analysis more computationally tractable.

We are, in any case, interested in the analysis of longer n -grams. The classification results for n -grams of lengths between three and five, reported in Table 8.7, provide evidence of the usefulness of this simple ranking approach based on filtering: the n -gram representation of the AA alphabet retains an accuracy of 0.92 with 268 attributes, while the n -gram representation of the SEZ alphabet achieves the same accuracy with 331.

The subsequent second-step, SVM-based FS process, starting from the optimal t-test selection was quite successful at reducing the number of attributes, while maintaining an accuracy of approximately 0.91 in the case of the AA alphabet for 45 features (a 83% reduction of the dimensionality) and a very reasonable 0.90 in the case of SEZ for 43 features (a 87% reduction), as seen in Table 8.8. In the case of the AA alphabet, the algorithm selects 6 *contiguous* and 39 *rigid gap n*-grams. For the SEZ alphabet, the 43 *n*-grams include 12 *contiguous* and 31 *rigid gap* ones. Tables 8.9 and 8.10 lists all these *n*-grams in the order they were selected by the SFFS procedure.

This list should be the starting point for proteomics experts to investigate the involvement of specific *n*-grams in structural and functional roles of the receptor. For class C GPCRS, this entails investigating motifs potentially related to the orthosteric site at the VFT, that is, the binding site of a ligand. The standing hypothesis for our study is that the *n*-grams shown to have the ability to discriminate between class C subtypes might be related to these binding sites, because the latter are meant to be subtype-specific in as much as each subtype binds to different ligands.

Note that we have not only provided a selected list of *n*-grams with the ability to discriminate the most between class C GPCR subtypes, but also an explicit ranking of relevance for these *n*-grams that experts can resort to. For obvious space limitations, we only show in some detail the three *n*-grams from each alphabet at the top of this ranking.

In the case of the AA alphabet, we consider the *rigid gap n*-grams WXXW (which is significant in 18 t-tests) and PXXFR (significant in 16 t-tests) and the *contiguous* YGR (significant in 17 t-tests). Figure 8.2.5 shows the corresponding relative frequencies per subtype as boxplot diagrams.

Figure 8.2.6 shows the corresponding boxplots for the three most discriminant *n*-grams from the SEZ alphabet. They are WXXW, G[DE]X[RKH] and [ST]XX[QN]ST, all of which are significant in 16 tests.

The AA *n*-grams discrimination capabilities seem to be mainly based on their existence, or lack of it, in sequences of different subtypes. This is consistent with the restrictive idea that a *signature* motif should be characterized as one that matches all the sequences of a given family and no sequence outside this family [15]. WXXW seems to be mostly absent in two of the main subtypes, namely mG and GB, whereas PXXFR appearance seems mostly restricted to GB and Ph, and YGR restricted to Ta. Note also that the grouping of some frequency values for some subtypes beyond the main quartiles of the boxplots is a hint to the existence of grouping structure within subtypes. For instance, YGR seems to be present with very specific frequencies not only in Ta, but also in small subgroups of mG

Table 8.9.: Lists of n -grams, from the AA alphabet, (ranked by relevance according to the sequential forward feature selection procedure for SVM classifiers. For each n -gram, the ranking order ($\#$), the symbolic subsequence(see Table 3.3.2), where X is the wildcard residue in *rigid gap* n -grams, and the number of binary classifiers in which the n -gram was found to be significantly discriminant (SIGN), are displayed.

$\#$	n -gram	SIGN	$\#$	n -gram	SIGN
1	WXW	18	24	YXXXY	16
2	PXXFR	16	25	CXEXC	16
3	YGR	17	26	VXXLL	16
4	WXWXG	17	27	SNXXD	16
5	CIA	16	28	SXKXQ	16
6	YXI	16	29	CXDG	17
7	AXXL	16	30	IXR	17
8	TGXE	19	31	WXXXL	16
9	GXXG	16	32	AWXXS	16
10	GEXXN	17	33	AXXSS	16
11	DCXXG	16	34	PGXXK	16
12	FPXH	16	35	GXRK	16
13	PNXXL	18	36	PNXT	16
14	WXL	17	37	VXCXD	16
15	QXMXF	16	38	GXXY	19
16	CXG	17	39	DCLP	16
17	IPG	16	40	GXCXA	16
18	HXXF	17	41	IXWH	16
19	CXXGT	17	42	CXXGT	17
20	YXKXG	17	43	CXAXS	16
21	DYG	16	44	YXD	16
22	PXIXY	16	45	VVFS	16
23	WXXV	16			

Table 8.10.: Lists of n -grams, from the SEZ alphabet, ranked by relevance according to the sequential forward feature selection procedure for SVM classifiers. For each n -gram, the ranking order ($\#$), the symbolic subsequence (see Table 3.4), where X is the wildcard residue in *rigid gap* n -grams, and the number of binary classifiers in which the n -gram was found to be significantly discriminant (SIGN), are displayed.

$\#$	n -gram	SIGN	$\#$	n -gram	SIGN
1	WXXW	16	23	[IVLM]XW	16
2	G[DE]X[RKH]	16	24	AXXX[ST]	16
3	[ST]XX[QN][ST]	16	25	C[RKH]XG	17
4	GXCC	16	26	[IVLM][IVLM]XW	16
5	CX[IVLM]	16	27	CXAX[RKH]	16
6	[QN]XWG	16	28	[QN]XGX[QN]	16
7	[ST][QN]A[RKH][IVLM]	17	29	[IVLM]XC[QN]	16
8	W[QN]X[QN]	18	30	W[ST]XX[IVLM]	16
9	[ST][QN][RKH][ST]	16	31	WX[RKH]W	16
10	PPX[ST]	17	32	W[QN]P	16
11	W[IVLM][QN][RKH][DE]	16	33	[DE]CXXC	17
12	[IVLM][IVLM][IVLM][ST]W	17	34	[QN]CC	16
13	[QN]X[QN]XW	16	35	[ST]XWW	16
14	[QN]GW[QN]	16	36	[ST]X[ST]X[QN]	16
15	[QN]X[IVLM]XC	16	37	[QN][QN]XX[ST]	16
16	[IVLM]GXXC	16	38	GXC[RKH]	16
17	[ST][QN]W[QN]	16	39	W[RKH]X[IVLM]	16
18	[IVLM]X[ST]XC	16	40	[QN][RKH][ST][RKH][IVLM]	16
19	[RKH]WX[IVLM]	19	41	[ST]X[RKH][ST]	16
20	[QN][ST]W	16	42	[RKH]XGXA	16
21	[ST][DE][ST]	16	43	[QN]XWX[ST]	16
22	PX[DE][ST]	16			

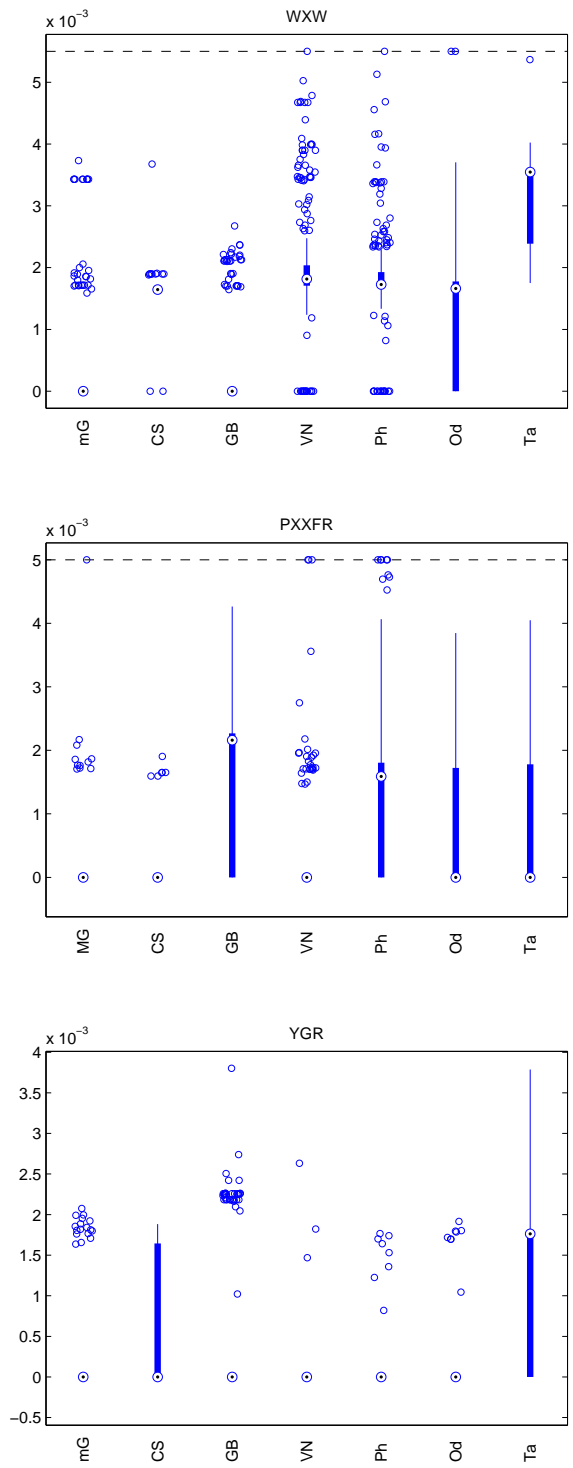


Figure 8.2.5.: Frequencies boxplots of the three n -grams of the AA alphabet ranked as the most discriminative in the classification of the 7 class C GPCR subtypes.

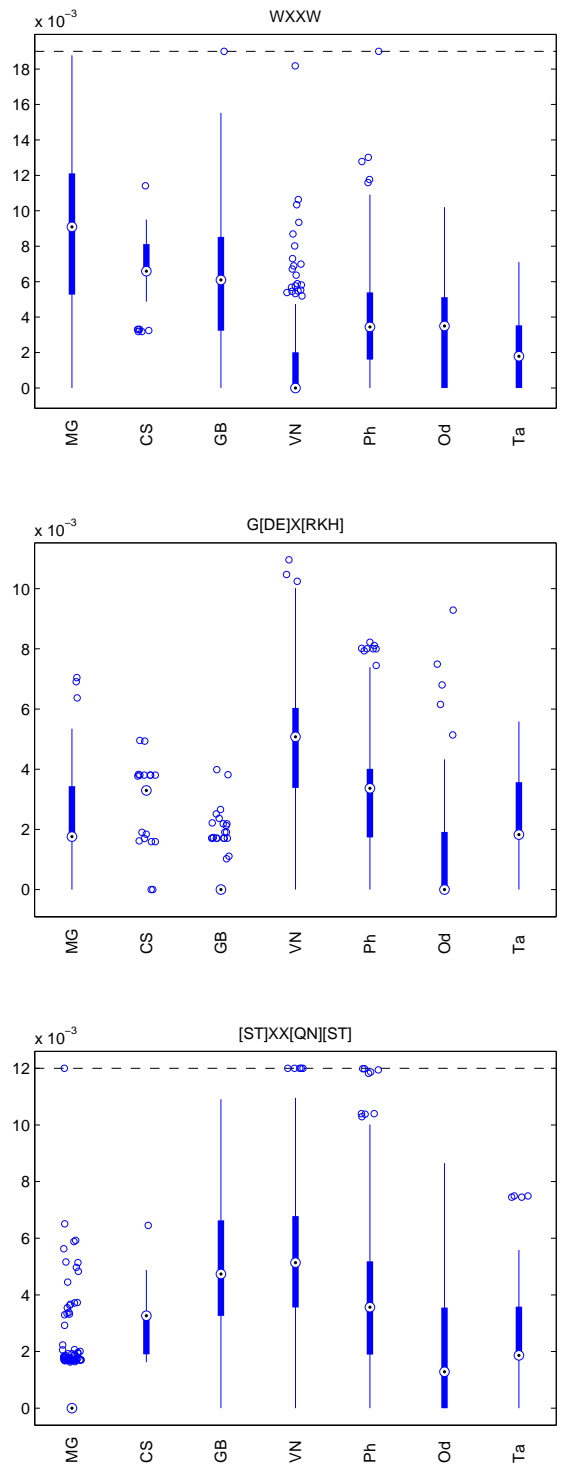


Figure 8.2.6.: Frequencies boxplots of the three n -grams of the SEZ alphabet ranked as the most discriminative in the classification of the 7 class C GPCR subtypes.

and GB.

The SEZ alphabet discrimination capabilities, instead, seem to be rather more subtle, as they are less based on the lack of a given n -gram than on a more gradual differentiation of the range of their frequencies. This a somehow natural consequence of their flexibility of sequence instantiation, resulting from the less granular use of the AA alphabet. WXXW seems very frequent in mG but infrequent in VN, Ph, Od and Ta. Instead, G[DE]X[RKH] is most frequent in VN and least in GB and Od, while [ST]XX[QN][ST] is mostly absent from mG (again with the exception of an eccentric but tight subgroup), but most frequent in GB and VN.

Conclusion

In this experiment, we have analyzed class C AA primary sequences from their *contiguous* and *rigid gap* n -gram frequencies, using a combination of FS and classification. This analysis involved class C subtype discrimination and aimed at identifying those n -grams most relevant to such task as candidate signature motifs. Motif over-representation in the sequence maybe the result of evolutionary preservation, which might be a lead to potential structural and functional roles. The selected discriminant n -grams may be related to the orthosteric sites at the VFT of the N-Terminal domain, given that these sites bind to different ligands for different subtypes and are thus subtype-specific.

Our previous research, using the frequencies of n -grams of length up to three obtained from the complete sequences, reported class C subtype classification accuracies that have been matched in the current study using the frequencies of a parsimonious selection of n -grams of length up to five obtained from just the N-terminal domain. Such results reinforce the interest of this extracellular domain in class C GPCR functional investigation. Of note also that the list of relatively long selected n -grams should be more effective than shorter ones as the starting point for proteomics experts to investigate motifs potentially related to the orthosteric site of the VFT, an investigation with clear potential in pharmacological research.

8.3. Feature selection used for the identification of subtype characteristic n -grams

In the current experiment, the results of the previous experiments are extended by investigating a different classification methodology: we switch from the subtype-*vs*-subtype classification employed in [12, 11] to a subtype-*vs*-all the rest of subtypes

procedure. That is, we shift towards the selection of those sequence motifs that distinguish each class C subtype from the rest.

8.3.1. Experiments on the N-terminus with a two-stage Feature Selection

Experimental settings

In this experiment we analyze the N-terminus segment of the sequences of the 2011 version dataset using a one-vs-rest classification methodology and a two stage dimensionality reduction approach combining a t-test filter and a SFFS as previously used in [11, 5]. Note that the dataset comprises the 1,252 sequences with a full 7TM structure as explained in Section 7.2.2 (See Table 7.1).

The experiments on the N-terminus comprise the following steps:

1. Data preprocessing: The primary AA sequences of the N-terminus are transformed for the AA alphabet to contiguous and rigid gap n -gram frequencies of lengths three to five. As in [12], the relative frequencies of occurrence of the n -grams (real-valued quantities), are employed for the classification experiments.
2. t-Test Filtering and SFFS: Given a feature space built by contiguous and rigid gap n -grams (of length three to five) the two step dimensionality reduction approach is applied:
 - a) As a first stage a t-test filter is applied, which due to the very large number of features performs a crude sorting out. It consists in two-sample t-tests between receptor subtypes for feature filtering, with 0.01 significance level. In previous research [5], a subtype-*vs*-subtype classification setting was employed, where the t-tests were run for the 21 subtype-*vs*-subtype possible combinations. Instead, a subtype-*vs*-rest of data scheme was used in the current experiment, so that a single t-test is run for each class C subtype. As a result, the t-test is meant to discern whether a feature is significantly discriminating a given subtype from the rest of the sequences. Only n -grams with a non-zero mean value for the specific subtype are considered, as we are interested in subtype-specific n -grams, i.e. n -grams that are actually present in the subtype.
 - b) The second dimensionality reduction stage involves a SFFS algorithm with identic settings as those of the previous FS experiments (See section 8.2.1).

- c) Assessment of the selection of features of the SFFS algorithm: The result of our FS approach is the subset of n -gram frequencies that are the outcome of this second-stage SFFS process. Therefore, we repeat the forward selection algorithm 10 times for each class C subtype and evaluate it by counting how many times each feature was selected. Furthermore, we take into account in which position an attribute was selected during the sequential selection (whether it was earlier or later in the process) and establish a positional ranking. More precisely, features are weighted according to the position in which they were selected, i.e., we count how many times a given feature was selected and weight the selection according to whether it was selected first (5 points), second (4 p.), third (3 p.), fourth (2 p.) or subsequently (1 p.).
3. We also explore in which position of the N-terminus the selected n -gram is located, either in the main part of the sequence or in the CRD. This distinction is important, given that the N-terminus attributes different functionalities to the VFT (built by the AAs from the beginning and middle part of the sequence) and the CRD (located at the end).

The classification performance of the models constructed from the different feature sets is again measured using the same CV approach and metrics as those of the previous FS experiments (See section 8.2.1 for a detailed explanation).

Results

t-Test Filtering In our experiments, we built SVM-based classification models using the n -grams from the N-terminal domain of the available sequences using a subtype-*vs*-rest of data setting instead of a subtype-*vs*-subtype approach as in previous research [12]. In more detail, we used *rigid gap motifs* (n -grams) of lengths three to five.

t-Test filtering was employed to generate a first crude ranking of features. In the current subtype-*vs*-rest of data approach, we show the SVM classification results per subtype using the reduced attribute set yielded by the t-tests in Table 8.11. Several standard classification metrics are reported, including Accuracy, MCC, F-measure, precision and recall. These results show that the t-test reduction step yields a selected subset that allows the SVM algorithm to achieve fairly good results for subtypes mG, CS, GB and Ta. Results for subtypes Vn, Ph, and specially Od are less accurate, which is consistent with results of previous studies [3, 10].

Sequential Forward Feature Selection The t-Test filtering method found feature subsets with high classification accuracy. Their dimensionality is still quite

Table 8.11.: Subtype classification results after t-test selection. $T - D$ is the size of a feature set yielded by the t-test reduction step and Accu stands for classification accuracy (ratio of correctly classified sequences), MCC for the Mathew correlation coefficient, F-Meas for F-measure, Prec for precision and Rec for recall.

SUBTYPE	T-D	Accu	MCC	F-Meas	Prec	Rec
mG	588	0.98	0.94	0.95	0.97	0.94
CS	918	0.99	0.94	0.94	1.0	0.89
GB	483	0.99	0.96	0.97	1.0	0.94
Vn	352	0.95	0.86	0.89	0.93	0.86
Ph	299	0.95	0.87	0.90	0.95	0.86
Od	329	0.95	0.52	0.52	0.76	0.4
Ta	464	0.99	0.99	0.99	1.0	0.98

high, though, which is why we applied the more nuanced second step of SVM-based SFFS. Table 8.12 summarizes the subtype classification results after applying the SFFS algorithm on the t-test reduced feature set. These results show that the algorithm has successfully reduced the number of features achieving the same or even better classification.

Table 8.12.: Subtype classification results after Forward Selection, where $T - D$ and $FW - D$ denote the size of a feature set after t-test and Forward Selection, respectively.

SUBCLASS	T-D	FW-D	Accu	MCC	F-Meas	Prec	Rec
mG	588	4	0.98	0.97	0.97	0.97	0.97
CS	918	6	0.99	0.96	0.96	1.0	0.93
GB	483	6	0.99	0.96	0.97	0.99	0.95
VN	352	25	0.93	0.8	0.84	0.9	0.79
Ph	299	24	0.93	0.83	0.87	0.88	0.86
Od	329	12	0.96	0.61	0.58	0.94	0.42
Ta	464	5	0.99	0.97	0.97	0.98	0.97

Comparison with subtype-*vs*-subtype classification results We now compare the classification results obtained with the subtype-*vs*-rest approach with those obtained in [12] using the subtype-*vs*-subtype approach (Table 8.13).

Table 8.13.: Subclass classification results using the subtype-*vs*-subtype approach after Forward Selection. $T - D$ and $FW - D$ denote the size of a feature set after t-test and Forward Selection respectively.

SUBCLASS	T-D	FW-D	MCC	F-Meas	Prec	Rec
mG	268	82	0.94	0.95	0.95	0.96
CS	268	82	0.94	0.94	1.0	0.89
GB	268	82	0.95	0.96	0.96	0.96
VN	268	82	0.88	0.91	0.91	0.91
Ph	268	82	0.87	0.90	0.90	0.90
Od	268	82	0.76	0.76	0.80	0.78
Ta	268	82	1.0	1.0	1.0	1.0

The per-subtype classification results reported in this research using the subtype-*vs*-rest approach are better than those obtained by the multi-class classifier using subtype-*vs*-subtype for subtypes mG, CS and GB; roughly similar for Ta and worse for Vn, Ph and Od. Note though that the n -gram selection is far more parsimonious than the one reported in [12](Table 8.13). For instance, an MCC value of 0.97 for mG is reported in Table 8.12 with only 4 n -grams, as compared with an MCC of 0.94 with 82 n -grams in [12](Table 8.13). Therefore, we focus only on the good performing subclasses mG, CS, GB and Ta on the two-step FS process of the algorithm.

Assessment of attribute selection In the experiment reported here, we focus further on the evaluation of the attribute selection. Therefore the SFFS process was repeated 10 times on the reduced t-test feature set evaluating the selection considering the number of times (selection count) and in which position the attribute was selected by the algorithm (position weighted selection). In the following, we show the results of the assessment of attribute selection for mG, CS, GB and Ta. For each subtype, we show the position-weighted selection graphically. Figure 8.3.1, for instance, shows the position-weighted evaluation of n -gram selection for mG and CS.

In the 10-time run of the algorithm a set of 2 to 6 attributes were selected for mG achieving similar test results as reported in section 8.3.1. The assessment of

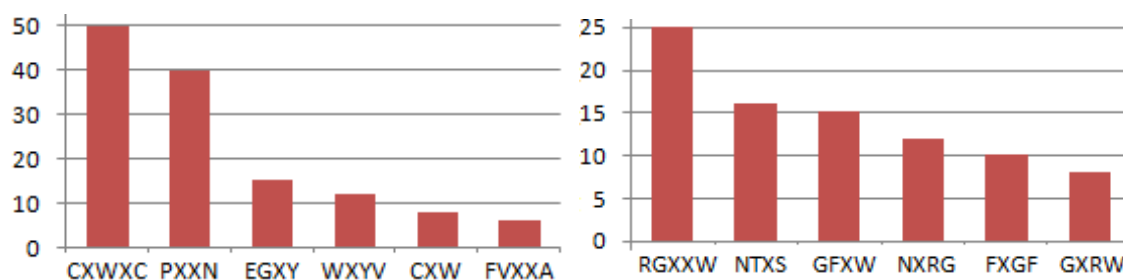


Figure 8.3.1.: Position-weighted selection of n -grams for mG (left) and CS (right).

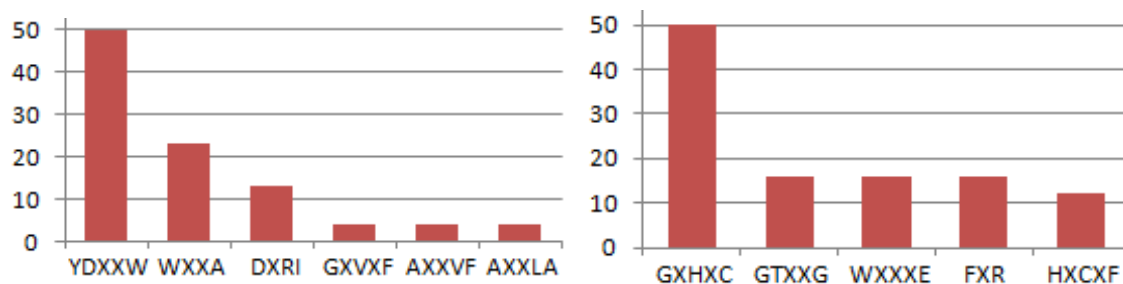


Figure 8.3.2.: Position-weighted selection of n -grams for GB (left) and Ta (right).

the attribute selection graphically shown in Figure 8.3.1 (left) depicts that two n -grams were most often selected. n -Grams CXWXC and PXXN should be analyzed in further detail as they have a high selection score. Both were selected in each of the 10 runs as the first and second attributes, respectively. n -Grams EGXY and WXYV also achieve a high score as they were selected frequently (5 and 4 times) as third attributes.

For CS, the test run yields a reduced feature set between 3 and 7 attributes. The selection assessment shows that n -grams RGXXW, NTXS, GFXW, NXRG, FXGF and GXRW were frequently selected. All these n -grams obtained a high score in the position-weighted selection as they were always between the first three selected attributes.

The corresponding test run for GB yielded data sets with 5 to 10 attributes. The attribute selection evaluation shows that n -grams YDXXW, WXXA, DXRI were frequently selected. The first one is selected in every run as first attribute. The latter two were selected in 6 and 5 runs, respectively, and always as second or third attributes. Also, n -grams GXVXF, AXXVF and AXXLA could be seen as characteristic, as they were selected in some runs as second attribute.

For Ta, the test yielded feature sets with between 4 and 6 attributes. The assessment shows some frequently selected n -grams of interest: GXHXC was selected in

each run as first attribute, while GTXXG, WXXE, HXCXF and FXR were selected frequently as second or third attributes. Table 8.14 summarizes all n -grams selected in the 10 runs.

Table 8.14.: Selected n -grams for each subtype.

mG		CS		GB		Ta	
CXWXC	-	RGXXW	PXTR	YDXXW	KXXP	GXHXC	WXWI
PXXN	-	NTXS	YXA	WXXA	PXDXT	GTXXG	-
EGXY	-	GFXW	FWXXE	DXRI	DAXW	WXXE	-
WXYV	-	NXRG	YAXS	GXVXF	EGXXG	HXCXF	-
CXW	-	FXGF	WXAS	AXXVF	YXAXW	FXR	-
FVXXA	-	GXRW	PXXCF	AXXLA	LAXN	VXGP	-
RXEXM	-	LXDXH	WXTS	GXV	GXSXP	QXMR	-
DAXXA	-	KVXP	LXAXP	QXL	FXRT	AXXGP	-
YEYE	-	AXEXW	-	IXE	DXN	NNXS	-
-	-	PXYF	-	MXG	-	DXCS	-

Detailed analysis of selected n -grams Table 8.15 shows an analysis of the n -grams detected as most important in the FS assessment. For each n -gram, the subtype specific location on the N-terminus is reported.

The pattern CXWXC is characteristic of mG sequences as it appears at 269 (tc=269) out of a total of 282 mG sequences (nc=282), while it appears only at 6 sequences of the remaining subtypes (to=6). This n -gram is located at the end of the sequence and is part of the CRD. PXXN (tc=243) is also found to appear frequently at mG. It nevertheless corresponds to different n -grams at different locations. Two examples of the existence of different n -grams matching the given pattern are PYNN, which appears 37 times at the middle of the sequence (tc=37) and PN[PIVSNHGAMKE]N, which appears 67 times in the CRD (tc=67). The n -gram EGXY (tc=175) was also found to be characteristic of subtype mG and corresponds mainly to n -gram EG[SDN]Y (tc=168). Its occurrences are in the main part of the N-terminus.

The n -gram RGXXW (tc=30) was detected as characteristic for CS (nc=45). This n -gram corresponds entirely to the n -gram RGFRW (tc=30), located at the main part of the sequence. The pattern NTXS (tc=37) represents two frequent n -grams in CS: The frequent n -gram NTVS, which appears as part of the longer n -gram

Table 8.15.: Location of n -grams. nc denotes the number of instances of the subtype. tc and to denote, in turn, the number of instances of this subtype, or the remaining subtypes, that match the n -gram.

N	N-gram	Subtype	nc	tc	to	Location
1	CXWXC	mG	282	269	6	CRD
2	PXXN	mG	282	243	708	-
3	EGXY	mG	282	175	3	main
4	WXYV	mG	282	216	9	main
5	RGXXW	CS	45	30	18	main
6	NTXS	CS	45	37	104	main
7	GFXW	CS	45	31	11	main
8	YDXXW	GB	156	140	5	main
9	WXXA	GB	156	148	485	main
10	DXRI	GB	156	131	75	main
11	GXVXF	GB	156	135	227	main
12	GXHXC	Ta	62	56	18	CRD
13	GTXXG	Ta	63	58	104	main
14	WXXXE	Ta	63	58	582	main
15	HXCXF	Ta	63	46	13	CRD
16	FXR	Ta	63	63	988	main

DTCNTVS in CS subtype ($tc=20$). The other frequent n -gram is NTES ($tc=16$), which forms part of n -gram [TS][CIL]FWNTES ($tc=16$). Both patterns are located at the main part of the sequence. The pattern GFXW ($tc=31$) corresponds entirely to the n -gram GFRW, which is part of the longer n -gram F[RL]GFRW ($tc=31$) and is located in the main part of the sequence.

The GB characteristic n -gram YDXXXW ($tc=140$, $nc=156$) corresponds mainly to the n -gram YD[GA][IV]W ($tc=131$), which forms part of the longer n -grams [PT][LFY]AYD[GA][IV]W ($tc=54$) and G[YF][TAS]YD[GA][IV]W ($tc=50$). An analysis of the location is unnecessary as GB sequences lack the CRD. WXXA ($tc=148$) corresponds mainly to WA[AEILMFTSV]A ($tc=84$) and WV[TILV]A ($tc=27$) and WS[TILVSA]A ($tc=34$). Pattern DXRI ($tc=131$) corresponds to pattern D[TIVAS]RI ($tc=131$). Furthermore, the more specific n -grams DARI and DVRI are very characteristic as each appears 60 and 49 times, respectively. Pat-

tern GXVXF (tc=131) is characteristic for GB as it appears in 131 sequences. The most frequent n -grams described by this pattern are GPV[AEGSR]F (tc=48), G[QH]VVF (tc=47) and G[ARFVY]VAF (tc=22).

The Ta characteristic pattern GXHXC (tc=56, nc=62) corresponds mainly to the n -gram G[QMSILDFV]H[ETHIQVLSK]C (tc=47), which is located at the beginning of the CRD. The pattern GTXXG (tc=58) corresponds to the n -gram GT[VF][IVL]G (tc=57) located at the main part of the sequence. WXXXE (tc=58) corresponds to the n -grams W[IVLT][IAG][TGS]E (tc=52) located at the main part of the sequence, while HXCXF (tc=46) corresponds to the n -gram H[ESTKPVIHQ]CCF (tc=46), located at the beginning of the CRD. Pattern FXR (tc=63) corresponds mainly to the longer n -gram [FY]PSF[YFLVM]R (tc=41) with occurrences in the main part of the sequence.

Comparison with known motifs This section presents a comparison of the n -grams listed in Table 8.14 with the known GPCR motifs to be found in the PRINTS-S¹ database, which constitutes a compendium for protein fingerprints linked with Swiss-Prot and TrEMBL, in turn important protein knowledgebases. The comparison is with the mG GPCR signatures (Identifier:GPCRMGR), the Extracellular CS receptor signatures (Identifier: CASENSINGR) and GB receptor signatures (Identifier: GABABRECEPTR). These signatures are derived from smaller sets of not necessarily coinciding sequences to those of GPCRdb.

Comparison to the GPCRMGR signatures shows that the mG specific n -gram EG[SDN] is part of the *Motif 7* located in the N-terminus. Another coincidence is found for the n -gram WXYV, which corresponds mainly to n -gram W[NT]YF. This n -gram forms part of *Motif 6*. Another known motif is the specific n -gram RLEAM (detected as RXEXM), part of *Motif 2*.

Comparison with CASENSINGR signatures shows that the n -grams found as subtype-specific do not form part of any known motif from CASENSINGR.

In turn, comparison of the n -grams found as characteristic for GB shows some coincidences with the known GABABRECEPTR signatures. The 13 elements comprising *Motif 6* match a long list of n -grams found by the sequential forward selection algorithm: The n -gram YD[GA][IV]W was found as characteristic and forms at the same time part of the longer n -gram [PF][LFY]AYD[GA][IV]W. A comparison shows that pattern AYD[GA][IV]W is part of *Motif 6*. Also, patterns YXAXW and DAXW describe this same subsequence. The n -gram WA[AEILMFTSV]A and its more specific n -grams WALA and WVIA constitute part of this motif as well. The n -gram AXXLA matches the more specific n -gram ALALA, also part of

¹<http://www.bioinf.manchester.ac.uk/dbbrowser/sprint/>

the known motif. The n -gram LAXN matches the end of the motif by representing the LALN pattern. The pattern FXRT matches n -gram FFRT, which is part of the 20-element long *Motif 3*. The n -gram D[TIVAS]RI represents the n -grams DARI and DVRI, part of the 24 elements comprising *Motif 4*.

Discussion

The two-step FS process using the subtype-*vs*-rest approach has successfully yielded very parsimonious subsets of attributes that provide similar or better classification results for subtypes mG, CS, GB and Ta to those obtained using the subtype-*vs*-subtype approach (See sections 8.3.1 and 8.3.1). Table 8.12 reports the SVM classification results for the reduced feature sets, where a MCC between 0.96-0.97 is achieved with feature sets of 4-6 attributes. The finding of such reduced feature sets provides support to the hypothesis of the existence of subtype-specific motifs.

The more detailed analysis of the FS process, aiming to find out whether the selected n -grams are characteristic for the particular subtype, was accomplished by repeating the SFFS process several times and evaluating the selected features according to their frequency of selection and position. The resulting position-weighted selection score has revealed a small subset of n -grams that are consistently selected for a given subtype by the algorithm (See section 8.3.1).

A comparison with known motifs from the PRINTS-S database has confirmed that several of the detected n -grams from subtypes mG and GB are part of known motifs in this database. This finding partially confirms the potential of the proposed approach for the selection of subtype-characterizing n -grams and encourages further research investigating the newly detected motifs from a proteomics viewpoint.

Conclusion

The results reported in this section reinforce the idea that ML methods are useful tools for knowledge extraction from GPCR data, especially if embedded in well-principled data dimensionality reduction procedures. We have shown that, using such methods, several parsimonious subsets of sequence n -grams, understood as receptor motifs, can successfully discriminate class C subtypes. Furthermore, several such n -grams match known motifs reported in standard curated protein databases.

The identification of these subtype-specific motifs should be the starting point for the investigation of their role in receptor functionality, as they could be related to the orthosteric sites at the VFT in the investigated extra-cellular N-Terminus. This may provide at least partial foundations to pharmacological research, as 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 [130].

8.3.2. Experiments on the N-terminus using χ^2 Filter selection

In this experiment we apply a χ^2 filter for FS on the n -gram frequency representations of the data. This univariate metrics measures how good a feature distinguishes the instances of a subtype from the instances of the rest of subtypes. The purpose of the present experiment is to systematically assess the FS of the greedy approach using a two step FS technique of the previous experiment on the N-terminus in order to verify in which degree the uppermost subtype specific n -grams have been selected. Moreover we want to compare the two stage FS approach with a FS technique that is able to operate on vast feature spaces, because the use of a forward selection algorithms is limited by the size of the feature space.

Experimental settings

This experiment focuses on the analysis of the N-terminus domain of the 2011 and September 2016 dataset version. Remember that these datasets comprise respectively 1,252 and 922 sequences with a full 7TM structure as explained in section 7.2 (See Table 7.1 and 7.11).

The purpose of the present experiment is to systematically assess the FS done by the two step FS technique of the previous experiment (See 8.3.1) in order to verify in which degree the most subtype specific n -grams have been selected. For this end we carry out the following steps:

1. χ^2 Filter approach: We calculate the χ^2 measure for the n -grams of each subtype and select the n features with highest χ^2 value. For those n -grams we report also the number of instances of this subtype (tc) and the number of instances of the remaining subtypes (to), that match the n -gram.
2. We compare the n -gram selection between the greedy approach of the previous experiment and the χ^2 selection as well as between different versions of the dataset.
3. We assess the internal data quality of the χ^2 selected n -gram subsets with supervised ML methods using the same CV approach and metrics as explained for the previous FS experiments (See section 8.2.1 for a detailed explanation).

Results

Overview of the n -gram feature space Table 8.16 shows the number of n -gram features for the 2011 and 2016 dataset using contiguous and rigid gap motifs of length 3 to 5. For each n -gram of a given length the contiguous and rigid gap n -grams are counted, for example the number of n -grams reported for 4-grams include all n -grams of length 4 with zero, one or two wildcards positions. Note that only n -grams which effectively appear in the data are considered. From the numbers reported in Table 8.16, we see how fast the feature space grows as the use of rigid gaps motifs enlarges even more the vast feature space of contiguous n -grams. The χ^2 filter approach calculates the χ^2 value for each of the n -grams of the feature space.

Table 8.16.: Number of n -grams of the 2011 and Sept. 2016 dataset (N-terminus only). Both contiguous and rigid gap motifs are counted for a n -gram of a given length.

	N-grams			
Dataset	3-grams	4-grams	5-grams	Total
2011	8,370	117,813	574,481	700,664
Sept. 2016	8,129	65,783	251,493	325,405

χ^2 **Filter** In this section we report for the subtypes mG, CS, GB and Ta the most discriminative n -grams according to the χ^2 filter selection. We report the 15 n -grams with the highest χ^2 value for each subtype both for the 2011 dataset and September 2016 one. In each table nc denotes the number of instances of the subtype, tc and to denote, in turn, the number of instances of this subtype, or the remaining subtypes, that match the n -gram.

Tables 8.17 and 8.18 show the n -grams selected according to their χ^2 value for subtype mG for each dataset. Tables 8.19 and 8.20 show the selection for subtype CS, Tables 8.21 and 8.22 for GB and Tables 8.23 and 8.24 for Ta.

Comparison of n -gram selection For subtype mG Tables 8.17 and 8.18 report the subset of n -grams selected by the χ^2 value. A comparison of these n -grams for the 2011 and September 2016 databases shows the following matches: CCWXC, CWXC, CXWXC, WXC, RNXWF, RNXW, NXWF, RNXXF and RNXWF appear in both versions of the dataset. Regarding the comparison between the FS of the greedy approach (See 8.14) and chi-square selection (Table 8.17) we observe

Table 8.17.: List of 15 n -grams selected by χ^2 selection for the Mg subtype (2011 dataset).

N-gram	nc	tc	to	χ^2	N-gram	nc	tc	to	χ^2
CXWXC	282	269	6	893.4	RNXW	282	210	6	690,65
CCWXC	282	268	6	889.95	WXYVS	282	203	2	687.49
CCW	282	268	16	839.97	WXYXS	282	215	11	682.772
CWXC	282	268	21	816.35	RXXWF	282	207	12	650.51
NXWF	282	219	6	721.57	SDXW	282	219	24	636.32
WXC	282	269	45	717.45	WFXE	282	190	4	632.26
RNXWF	282	206	1	703.169	WXXVS	282	209	25	598.21
WXYV	282	216	9	696.08					

Table 8.18.: List of 15 n -grams selected by χ^2 selection for the mg subtype (2016 dataset).

N-gram	nc	tc	to	χ^2	N-gram	nc	tc	to	χ^2
CXWXC	505	479	1	578.5	RYD	505	449	1	542,1
CWXC	505	479	2	574.9	NXWF	505	448	1	540.9
CXWXC	505	479	5	563.9	NXRN	505	441	3	525
CXXM	505	479	8	546.9	RNXXF	505	451	7	522.6
WXC	505	479	10	546	KXXFV	505	418	1	504.4
RXXWF	505	446	0	542.2	YXAXY	505	418	1	504.3
RNXWF	505	446	0	542.2	GRY	505	431	7	498.4
RNXW	505	449	1	542.1					

that n -grams CXWXC and WXYVS have been selected by both approaches. The n -gram CXWXC has the highest χ^2 value and was selected according to the FS assessment always as first feature, while n -gram WXYVS (with fourth position in the FS assesment) has the 10th-highest χ^2 value.

For subtype CS Tables 8.19 and 8.20 report the n -grams with highest χ^2 value. A comparison of these n -grams for the 2011 and September 2016 reveal no matches between the subsets. Comparing the FS of the greedy approach (See 8.14) and χ^2 selection (See Table 8.19) we observe that n -grams GFXW, NXRG, FXGF and GXRW of the greedy approach match respectively the n -grams GFRW, NFRGF,

Table 8.19.: List of 15 n -grams selected by χ^2 selection for the CS subtype (2011 dataset).

N-gram	nc	tc	to	χ^2	N-gram	nc	tc	to	χ^2
GTRKG	45	38	3	893.1	AADDD	45	32	1	830.4
GTXKG	45	38	5	891.9	GGXIG	45	38	8	828.8
TRKGI	45	35	2	884.3	TXKGI	45	35	5	812.7
GXRKG	45	38	6	869.9	RGFR	45	30	0	804.6
TRKXI	45	38	5	865.5	NFRGF	45	30	0	804.6
GTRXG	45	39	8	854.8	FRGFR	45	30	0	804.6
GFRW	45	31	0	831.5	RGFRW	45	30	0	804.6
GGTIG	45	33	2	830.8					

Table 8.20.: List of 15 n -grams selected by χ^2 selection for the CS subtype (2016 dataset).

N-gram	nc	tc	to	χ^2	N-gram	nc	tc	to	χ^2
IXXIE	103	91	3	952.2	MAXXI	103	82	5	830
WNWXG	103	84	1	901.6	WNXXG	103	84	7	828.3
KXIE	103	88	5	895.1	TAXXI	103	88	11	827.6
VIXVF	103	86	6	861.5	DDDXG	103	82	6	818.2
NWXG	103	84	5	851.6	WXGXI	103	85	9	816.8
VIVVF	103	85	6	850.7	VIVXF	103	86	11	806.2
VIVV	103	85	6	850.7	MIXXI	103	93	20	793
RXLN	103	86	8	838.7					

FRGF and GFRW, which are all features with high χ^2 value.

In reference to subtype GB, Tables 8.21 and 8.22 show the subset of n -grams selected by the χ^2 value. A comparison between the selected n -grams for the 2011 and 2016 dataset show the subsequent matches: YDXXW, DXRII, RIIXG, FCXXY are selected for both datasets. YDAXW selected for the 2011 dataset is very close to the YDGXW pattern of the 2016 dataset. Note that this similarity could not be random as the AAs Alanine (A) and Glycine (G) have common physicochemical properties (both are aliphatic, non polar and neutral in charge).

Table 8.21.: List of 15 n -grams selected by χ^2 selection for the GB subtype (2011 dataset).

N-gram	nc	tc	to	χ^2	N-gram	nc	tc	to	χ^2
YDXXW	156	140	5	940.2	GWY	156	82	4	541.7
DXRII	156	125	3	851.8	YDXIW	156	76	0	533.9
RIIXG	156	119	1	827.1	FCXXY	156	79	4	520.7
DXRI	156	130	29	700.1	WAXAL	156	77	3	514.9
YDAXW	156	84	0	590.15	RII	156	141	90	499.8
YXAXW	156	89	6	574.6	AXXVF	156	107	39	495.3
WXXAL	156	114	33	570.9	DXRXI	156	131	75	493.8
DAXW	156	87	5	568.6					

Table 8.22.: List of 15 n -grams selected by χ^2 selection for the GB subtype (2016 dataset).

N-gram	nc	tc	to	χ^2	N-gram	nc	tc	to	χ^2
YDXXW	88	85	0	1101.9	DGXW	88	73	2	916.2
DXRII	88	85	0	1101.9	WIXXG	88	74	3	914.7
RIIXG	88	84	0	1089	WIXPG	88	70	0	907.5
FCXXY	88	77	1	982.9	YXWII	88	70	0	907.5
YXWI	88	74	0	959.3	WIIP	88	70	0	907.5
YDGXW	88	73	0	946.3	WIXP	88	72	2	903.3
SKXHG	88	72	0	933.4	KXHG	88	72	2	903.3
YXGXW	88	73	1	931.1					

For subtype Ta, Tables 8.23 and 8.24 report the n -grams with highest χ^2 value. Comparing the selection between the two datasets we see that n -grams GXHXC and GDYXL are selected for both. Regarding the matches between the greedy approach and the χ^2 selection for the 2011 dataset we observe that the n -grams GXHXC and HXCXF have been selected for both approaches. Moreover the n -gram GTXXG, VXGP and AXXGP of the greedy approach describes the n -gram GTXLG, AVIGP and AXIGP or AVIGP, which are all features with high χ^2 value.

Table 8.23.: List of 15 n -grams selected by χ^2 selection for the Ta subtype (2011 dataset).

N-gram	nc	tc	to	χ^2	N-gram	nc	tc	to	χ^2
HXCCF	62	46	0	882.9	GTXLG	62	43	0	747.8
AVIGP	62	49	5	844.3	GDYXL	62	42	3	746.8
AVIXP	62	49	5	844.3	HXXCF	62	46	11	694.8
AXIGP	62	54	11	842.8	GTVXG	62	37	1	689.5
AVXGP	62	49	8	794.8	HXCXF	62	64	13	668.2
HXCXF	62	56	18	786.3	VWXAS	62	42	8	663.7
FLXPQ	62	43	2	784.8	LLXGL	62	41	9	630.6
HXCC	62	59	23	782.0					

Table 8.24.: List of 15 n -grams selected by χ^2 selection for the Ta subtype (2016 dataset).

N-gram	nc	tc	to	χ^2	N-gram	nc	tc	to	χ^2
VYXVA	226	200	2	841.8	CFXR	226	182	2	764.8
VYXV	226	200	2	841.8	PXQL	226	183	4	755.9
AVYXV	226	194	2	816.1	PXQLL	226	179	2	752
GXHXC	226	205	10	812	PWQLL	226	177	2	743.5
LHXXL	226	215	20	795.8	PWQL	226	178	3	741.1
VXEIN	226	184	1	780.1	GDYXL	226	173	1	733
WQL	226	187	4	773.1	VEEXN	226	172	1	728.8
WQLL	226	185	3	771					

Comparison of classification performance Previous research [8] revealed the existence of very reduced subsets with very high classification accuracy for subtypes mG, CS, GB and Ta. In this section we compare the classification performance of the χ^2 filter selected subsets with those of the previous experiments (for the 2011 dataset). Table 8.25 shows a comparison per-subtype of the classification results obtained with the two stage FS method and the χ^2 filter method by SVM. We report either the size of the feature set after Forward Selection (FW-D) or the size of the feature set selected according to the upmost χ^2 values.

Table 8.25.: Subtype classification results after feature selection, where $FW - D$ and χ^2 denote the size of a feature set after Forward Selection or χ^2 Filter selection, respectively.

SUBCLASS	FW-D	χ^2	Accu	MCC	F-Meas	Prec	Rec
mG	4	-	0.98	0.97	0.97	0.97	0.97
	-	5	0.982	0.949	0.96	0.971	0.95
	-	15	0.987	0.963	0.971	0.979	0.965
CS	6		0.99	0.96	0.96	1.0	0.93
		5	0.9912	0.869	0.869	0.935	0.82
		15	0.992	0.902	0.905	0.975	0.84
GB	6		0.99	0.96	0.97	0.99	0.95
		5	0.987	0.94	0.946	0.968	0.93
		15	0.993	0.961	0.971	0.969	0.974
Ta	5		0.99	0.97	0.97	0.98	0.97
		5	0.996	0.968	0.969	0.94	1.0
		15	1.0	1.0	1.0	1.0	1.0

Discussion

In this section we compared the FS of the two-step SFFS approach and χ^2 filter approach. Regarding the selection of subtype discriminating n -grams we have seen that the χ^2 filter approach provides a full overview of the most discriminating n -grams, which may include repetitions of similar n -gram patterns. The greedy SFFS approach instead only selects some of the most discriminative patterns avoiding similar patterns, what is consequence of the operating mode of the SFFS algorithm seeking for compact feature sets without redundancies.

Regarding the construction of classification models the greedy SFFS approach provided very reduced feature sets with highly accurate classification results. The χ^2 filter approach achieved similar results, but requiring a higher number of attributes in order to obtain equivalent classification results according to the MCC metric. An increase to a number of 15 features was necessary to equalize (in the case of GB) or even surpass (in the case of Ta) the results of the two-step FS subset. For subtypes mG and CS the classification performance was still slightly lower.

Conclusion

We carried out an analysis of FS with a χ^2 filter, which provided a broad overview about the set of the most discriminative n -grams for each subtype. The χ^2 filtering approach can be applied even with larger feature spaces than those of the current experiment, as its computational cost is linear on the size of the feature set. A comparison of the FS with a χ^2 filter and the two-stage SFFS approach of previous research revealed coincidences in some of the selected attributes, i.e. we confirmed the effectiveness for the greedy two-stage SFFS approach to select highly discriminative patterns. We also conclude that regarding the construction of robust classification models the two-stage SFFS approach provided the most reduced feature sets with high internal data quality, which was verified by very accurate classification results.

9. Analysis of 3-D crystal structures

9.1. Introduction

The investigation of protein functionality and signalling mechanisms is often based on the knowledge of crystal 3-D structures. In eukaryotic cell membrane proteins such as GPCRs, this knowledge is partial and fairly recent: The first GPCR crystal 3-D structure was fully-determined in 2000 [32] and over the last decade, the structures of some other GPCRs, most belonging to class A, have been solved [33]. In the case of class C GPCRs the information about full tertiary and quaternary structure is very limited, although recent impressive advances in the discovery of GPCR crystal structures [34, 35] were made.

In previous research we have systematically analyzed the primary sequences of class C GPCRs in search for subtype specific n -gram motifs. Applying different FS methods we have located sets of subtype specific n -grams for mG, CS, GB and Ta (See section 8.3).

In order to find out whether these short n -gram motifs may have a biological significance, for example related to the orthosteric binding site, we resort to the available 3-D crystal structures of class C GPCRs. The 3-dimensional structures may provide information about the ability of the constituent n -gram subsequences to have a given structural or biological function when analyzed by biochemical experts. We are also interested in analyzing whether the subtype specific n -grams detected by our approach are already known to have a structural or biological functionality and therefore we recollect known functional information about the crystalized sequences and analyze them with regard to each of the subtype specific n -grams.

9.2. Experiments with crystal structures of the N-terminus

In this section we present the experiments related to the crystal structures of the N-terminus of class C GPCR. In previous research [13] we have detected for

subtypes mG, Cs, GB and Ta a set of subtype specific n -grams discovered from the 2011 and September 2016 dataset (See Section 8.3). In this experiment we analyze the available crystal structures of the N-terminus of class C GPCRs with regard to these subtype specific n -grams. Despite the scarceness of full 3-D crystal structures of class C GPCRS, which comes from the difficulty to solve the transmembrane regions, the crystalization of the extracellular domain is quite feasible, therefore being several partial crystal structures available for the N-terminus. Note that the available crystal structures of the extracellular domain do not belong necessarily to the analyzed sequences of the GPCR dataset from GPCRdb, but given the very limited number of crystal structures of the sequences of this dataset, our analysis is extended to all available crystal structures of class C GPCRS.

9.2.1. Experimental settings

Methods Our experiment focuses on the analysis of the information related to known 3-D structures of class C GPCRS in regard to the subtype specific n -grams detected by our approach. The employed methodology comprises the following tasks:

1. Recollection of crystal structures: The 3-D structures from the Protein Data Bank (PDB) and its crystalized sequences are recollected. We search the crystal structures of the extracellular domain of class C GPCRS and the underlying crystalized protein sequence identified with a given Uniprot Accession ID. As GPCRS may be complexed with different substances or they may experience conformational changes in different activation states, it is frequent to find several crystal structures for a GPCR sequence which correspond to different states of the receptor. First we check whether the n -gram motifs are part of the crystalized sequence, what is a requisite to proceed with the deeper analysis explained in the next step.
2. Detailed analysis of the crystalized protein sequence in reference to a n -gram:
 - a) We determine the frequency of appearance of the n -gram in the sequence.
 - b) For each appearance of the n -gram pattern we document the position in the sequence and information whether the segment is part of a known secondary substructure (strand or helix).
 - c) By means of the graphical representation of the crystal structure with the tool UCSF Chimera¹, we determine whether the n -gram segment

¹<http://www.cgl.ucsf.edu/chimera>

pertains to a secondary structure (helix or strand) and whether it is located on the surface or in the inside of the receptor.

- d) Retrieval of known functionalities from Uniprot: We analyze whether the segment has an already known functionality, such as binding site, etc. or other structurally relevant information, for example whether it is part of a disulfide bond (DSB).

Tools In this section we describe the use of the external protein databases and protein structure visualization tool (UCSF Chimera) used in the aforementioned method. We illustrate the use of the PDB and UniprotKB database with an example:

Regarding the recollection of the 3-D structures, Figure 9.2.1 shows the PDB web interface with information for a 3-D structure in order to illustrate the retrieval of a crystal structure from the PDB repository. The screen shows the information related to the crystal structure '2E4Z', which corresponds to the ligand-binding region of the group III metabotropic glutamate receptor. This structural database provides moreover information about the 3-D structure, which can be visualized online or downloaded, information regarding the deposition and its authors and a link to the scientific publication related to the 3-D structure. A link to the UniprotKB with its Uniprot Accession Id is provided for access to the information about the protein with full detail.

Regarding the recollection of annotated functions and structural information from UniProtKB we show in Figure 9.2.2 the web interface with information for the sequence with Uniprot Accession Code P23385. This protein is characterized as Metabotropic glutamate receptor 1 obtained from a rat. In reference to the information available on UniprotKB we focus on the sections 'Functions' and 'PTM/Processing' to retrieve information about known functionalities. In this example the mG sequence has five known binding sites for Glutamate (as described in 'Functions') and several disulfide bonds (as described in 'PTM/Processing'). The section 'Structure' provides an overview of the related 3-D structures from PDB. In this example there are five crystal structures available, which are identified by its PDB entry code, and which interestingly comprise the same sequence segment, i.e. chains A/B from position 33 to 522 of the primary AA sequence. These five crystal structures belong to the receptor complexed with different substances, what may result in different states of the receptor. For example structure 1EWK corresponds to the receptor complexed with glutamate while structure 1EWS corresponds to a ligand free form and 1ISS captures the state of the receptor when complexed with an antagonist.

Regarding the visualization of the crystal structure we use the software UCSF

RCSB PDB 134656 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands **Go**

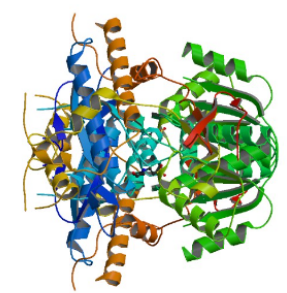
Advanced Search | Browse by Annotations | Search History (1) | Previous Results (5)

PDB-101 PDBe EMBL EMDataBank Bioinformatics Resource Project Worldwide Protein Data Bank

Take the RCSB PDB User Survey

Structure Summary **3D View** Annotations Sequence Sequence Similarity Structure Similarity Experiment Literature

Biological Assembly 1



2E4Z

Crystal structure of the ligand-binding region of the group III metabotropic glutamate receptor
DOI: 10.2210/pdb2e4z/pdb

Classification: **SIGNALING PROTEIN**
Deposited: 2006-12-17 Released: 2007-02-27
Deposition author(s): [Muto, T.](#), [Tsuchiya, D.](#), [Morikawa, K.](#), [Jingami, H.](#)
Organism: [Rattus norvegicus](#)
Expression System: Trichoplusia ni

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 3.3 Å
R-Value Free: 0.324
R-Value Work: 0.270

wwPDB Validation

Metric	Percentile Ranks	Value
R-facet		0.323
Clostrscore		11
RAMACHANDRAN outliers		1.99%
Steric strain outliers		25.7%
rotamers outliers		5.3%

View in 3D: NGL or JSmol (in Browser)

Standalone Viewers
[Simple Viewer](#) [Protein Workshop](#)
[Ligand Explorer](#) [Kiosk Viewer](#)

Protein Symmetry: Cyclic - C2 (View in 3D)
Protein Stoichiometry: Homo 2-mer - A2
Biological assembly 1 assigned by authors

Macromolecule Content
• Unique protein chains: 1

Literature [Download Primary Citation](#)

Structures of the extracellular regions of the group II/III metabotropic glutamate receptors
[Muto, T.](#), [Tsuchiya, D.](#), [Morikawa, K.](#), [Jingami, H.](#)
(2007) Proc.Natl.Acad.Sci.Usa **104**: 3759-3764
PubMed: [17360426](#) [Search on PubMed](#)
PubMedCentral: [PMC1820657](#)
DOI: [10.1073/pnas.0611577104](#)
Primary Citation of Related Structures: [2E4U](#) [2E4V](#) [2E4W](#) [2E4X](#) [2E4Y](#) [2E4Z](#)

PubMed Abstract:
Metabotropic glutamate receptors play major roles in the activation of excitatory synapses in the central nerve system. We determined the crystal structure of the entire extracellular region of the group II receptor and that of the ligand-binding region of the

Macromolecules

Classification: **SIGNALING PROTEIN** [Sequence Display for 2E4Z](#)

Total Structure Weight: 56226.11

Macromolecule Entities [Toggle Protein Feature View](#)

Molecule	Chains	Length	Organism	Details
Metabotropic glutamate receptor 7	A	501	Rattus norvegicus	Fragment: ligand-binding region, residues 33-521 Gene Name(s): Grm7 Gprc1g Mglur7

Protein Feature View - UniProtKB AC: [P35400](#) [UniProt](#) [Full Protein Feature View for P35400](#)

Figure 9.2.1.: Protein Data Bank Screen for structure 2E4Z.

UniProtKB - P23385 (GRM1_RAT)

[Basket](#)

[BLAST](#) [Align](#) [Format](#) [Add to basket](#) [History](#)

[Feedback](#) [Help video](#) [Other tutorials and videos](#)

Protein | **Metabotropic glutamate receptor 1**

Gene | **Grm1**

Organism | *Rattus norvegicus* (Rat)

Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level¹

Function¹

G-protein coupled receptor for glutamate. Ligand binding causes a conformation change that triggers signaling via guanine nucleotide-binding proteins (G proteins) and modulates the activity of down-stream effectors. Signaling activates a phosphatidylinositol-calcium second messenger system. May participate in the central action of glutamate in the CNS, such as long-term potentiation in the hippocampus and long-term depression in the cerebellum. [4 Publications](#)

Miscellaneous

Activated by quisqualate > glutamate > ibotenate > trans-1- aminocyclopentyl-1,3-dicarboxylate; inhibited by 2-amino-3-phosphonopropionate.

Sites

Feature key	Position (s)	Description	Actions	Graphical view	Length
Binding site ¹	74	Glutamate 2 Publications			1
Binding site ¹	165	Glutamate 2 Publications			1
Binding site ¹	236	Glutamate 2 Publications			1
Binding site ¹	318	Glutamate 2 Publications			1
Binding site ¹	409	Glutamate 2 Publications			1

PTM / Processing¹

Molecule processing

Feature key	Position (s)	Description	Actions	Graphical view	Length
Signal peptide ¹	1 – 18	Sequence analysis	Add BLAST		18
Chain ¹ PRO_0000012924	19 – 1199	Metabotropic glutamate receptor 1	Add BLAST		1181

Amino acid modifications

Feature key	Position (s)	Description	Actions	Graphical view	Length
Disulfide bond ¹	67 ↔ 109				
Glycosylation ¹	98	N-linked (GlcNAc...) asparagine 1 Publication			1
Disulfide bond ¹	140	Interchain			
Glycosylation ¹	223	N-linked (GlcNAc...) asparagine 1 Publication			1
Disulfide bond ¹	289 ↔ 291				
Disulfide bond ¹	378 ↔ 394				
Glycosylation ¹	397	N-linked (GlcNAc...) asparagine Sequence analysis			1
Disulfide bond ¹	432 ↔ 439				
Glycosylation ¹	515	N-linked (GlcNAc...) asparagine Sequence analysis			1
Disulfide bond ¹	657 ↔ 746	By similarity			
Modified residue ¹	853	Phosphoserine By similarity			1

Structure¹

Secondary structure

1

Legend: Helix Turn Beta strand PDB Structure known for this area

[Show more details](#)

3D structure databases

Select the link destinations:	PDB entry	Method	Resolution (Å)	Chain	Positions	PDBsum
<input checked="" type="radio"/> PDB ¹	1EWK	X-ray	2.20	A/B	33-522	[*]
<input type="radio"/> RCSB PDB ¹	1EWT	X-ray	3.70	A/B	33-522	[*]
<input type="radio"/> PDBj ¹	1EWV	X-ray	4.00	A/B	33-522	[*]
	1ISR	X-ray	4.00	A	33-522	[*]
	1ISS	X-ray	3.30	A/B	33-522	[*]
ProteinModelPortal ¹	P23385.					
SMR ¹	P23385.					
ModBase ¹	Search...					
MobiDB ¹	Search...					

Miscellaneous databases

EvolutionaryTrace¹ [P23385.](#)

Figure 9.2.2.: UniprotKB entry for protein P23385.

Chimera. Figure 9.2.3 shows the visualization of the two AA chains of the protein with their backbone structure. The structure at the left corresponds to the chain A (sequence with Uniprot ID Q9UBS5) and the right to chain B (sequence with Uniprot ID O75899). In chain A the n -gram ARKVF is highlighted in cyan, the n -gram YDAIW appears in green color, while the similar pattern YDGIW is highlighted in magenta in chain B. All three n -grams pertain to the secondary structure of a helix. Using a sphere presentation of the atoms a surface rendering of the receptor is done as shown in Figure 9.2.4. We appreciate in the upper structure (4MQE) that all three highlighted n -grams (ARKVF, YDAIW and YDGIW) are almost not visible, because they are located in the inside of the receptor. It is useful to evaluate the surface location as hint for the ability for ligand binding. An example for surface location is shown in the structure at the bottom of Figure 9.2.4. The pattern RNPWF of structure 1EWK has all five AAs located at the surface (see the green pattern highlighted in the structure at the bottom).

9.2.2. Results

Recollection of crystal structures

Table 9.1 summarizes the information about the crystal structures grouped by sequence segment for the mG subtype. There are often several structures for a sequence segment corresponding to different states of the receptor. For example sequences P23385 and P31422 have both five 3-D structures for the respective sequence segment. The sequences are described with their Uniprot Accession Id and the 3D structures use the PDB entry code for identification.

Table 9.2 shows the analysis of the crystalized sequences of mG (detailed in Table 9.1) with regard to the n -grams detected as frequent and subtype specific according to the results reported in the previous chapter (Tables 8.17 and 8.18). For each of the 14 crystallized sequence segments identified with M1-M14 we analyze whether one of the n -grams listed in the header of the table appears in the sequence. We report for each sequence segment the specific n -gram which was detected by the pattern. For example the group RNXWF, RNXW, NXWF, RXXWF and RNXXF match mostly the n -gram RNPWF, but for sequences M13 and M14, it matches the n -gram RNVWF or RNVNF. In the case of the short pattern GRY and RYD in some cases the pattern appears twice. In these cases the frequency is indicated between braces, i.e. RYD(2) denotes RYD appears two times in the sequence segment.

From the results reported in Table 9.2 we can see that the crystalized sequences M2, M3, M10 and M11 are not of further interest for us as they do not contain



Figure 9.2.3.: Visualization of chains A (left) and B (right) of structure 4MQE

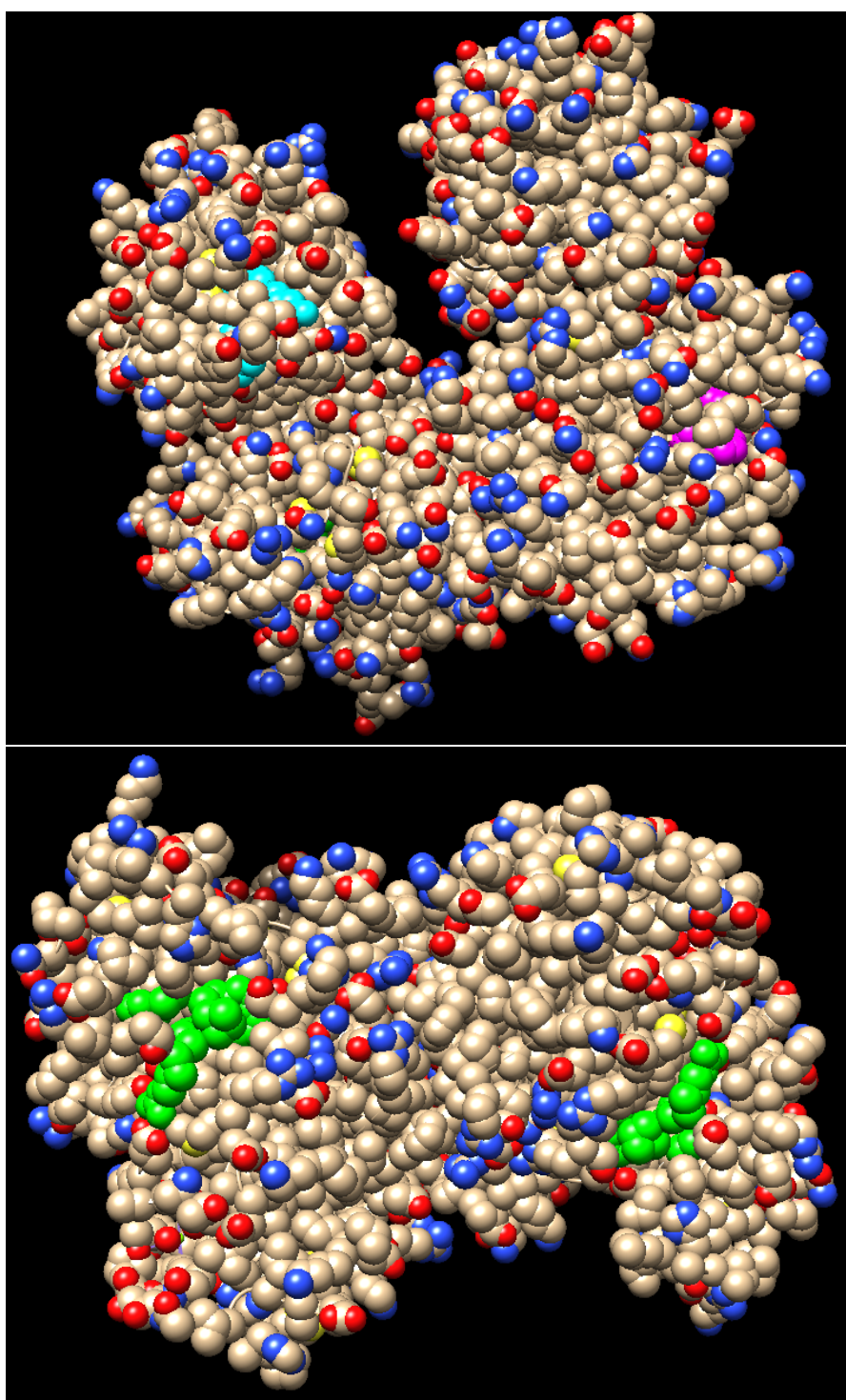


Figure 9.2.4.: Surface rendering of structures 4MQE (upper) and 1EWK (bottom).

any of the subtype specific n -grams of mG. We focus the detailed analysis of the other crystalized sequences in the following.

Table 9.3 summarizes the information about the crystal structures grouped by sequence segment for the CS subtype. The available 3-D structures relate both to the Uniprot sequence P41180, although they reference segments with slightly different lengths. The sequences are described with their Uniprot Accession Id and the 3D structures use the PDB entry code for identification.

Table 9.4 shows the analysis of the the crystalized sequences of CS detailed in Table 9.3 with regard to the n -grams detected as frequent and subtype specific for CS (Tables 8.19 and 8.20). For each of the two sequence segments identified with C1 or C2 we analyze whether one of the n -grams listed in the header of the table appears in the sequence. We also report for each sequence segment the specific n -gram which was detected by the pattern.

Table 9.5 summarizes the information about the crystal structures grouped by sequence segment for the GB subtype. Crystal structures 4MR7, 4MR8, 4MR9, 4MRM, 4MS1, 4MS3 and 4MS4 are both related to sequence Q9UBS5 (G2) and O75899 (G4), which are chain A and B of the protein respectively.

The sequences are described with their Uniprot Accession Id and the 3D structures use the PDB entry code for identification.

Table 9.6 shows the analysis of the crystalized sequences of GB detailed in Table 9.5 with regard to the n -grams detected as frequent and subtype specific for GB (according to the results reported in Tables 8.21 and 8.22). For each of the sequence segments identified with G1 - G4 we analyze whether one of the n -grams listed in the header of the table appears in the sequence. We report for each sequence segment the specific n -gram which was detected by the pattern. From the results we can see that only the crystalized sequences G2 and G4 are of interest for further analysis as they match with several n -grams. Sequence G5 corresponds to a short sequence not part of the N-terminus and is therefore not further analyzed.

A search at PDB revealed that for subtype Ta no crystal structures are available. In consequence no results can be reported for Ta in this experiment.

Detailed analysis of the n -grams

In this section we present the results of the detailed analysis of the crystal structures with reference to the subtype specific n -grams.

Crystal structures of Metabotropic Glutamate Receptor subtype 1 (related to sequence P23385)

In this section we review the crystal structures characterized as mG receptor of subtype 1, which comprises the sequence with UniprotKB Id P23385 as chain A and B. Table 9.7 shows an overview of the four different crystal structures, which are available for the sequences. A detailed analysis of the n -grams shows that the n -gram SDGW and KMGFV are already known as binding site for Glutamate. The other n -grams of Table 9.8 are not attributed currently a known functionality.

Crystal structures of Metabotropic Glutamate Receptor subtype 2 (related to sequence Q14416)

In this section we review the crystal structures characterized as mG receptor of subtype 2, which comprises the sequence with UniprotKB Id Q14416 as chain A and B. Table 9.9 shows an overview of the six different crystal structures, which are available for the sequences. A detailed analysis of the n -grams shows that the n -gram SDGW and KIMFV are already known as binding site for Glutamate. N-gram CDAM is involved in a DSB. The other n -grams of Table 9.10 are not attributed currently a known functionality.

Structure	Description
4XAQ	mGluR2 ECD and mGluR3 ECD with ligands
4XAS	MgluR2 EDC ligand complexed
5CNI	Mglu2 with Glutamate
5CNJ	Mglu2 with Glutamate analog
5KZN	Mglu2/3 Receptor Antagonist with Antidepressant Like Activity
5KZQ	Mglu receptor with Antagonist

Table 9.9.: Crystal structures for mG receptor subtype 2.

N-gram	Pos.	Binding Site	DSB	Structure
RYD	177-179	-	-	-
WTYVS	204-208	-	-	conn. helix to strand
SDGW	294-297	295, Glutamate	-	-
NSRN	339-342	-	-	-
RNPWF	341-345	-	-	partly helix
WFRE	344-347	-	-	helix
KIMFV	377-381	377, Glutamate	-	helix
CDAM	407-410	-	400-407	helix
GRY	451-453	-	-	partly strand
GRY	464-466	-	-	partly strand

Table 9.10.: Detailed analysis of functions and structures related to Q14416.

Crystal structures of Metabotropic Glutamate Receptor subtype 3 (related to sequence P31422) In this section we review the crystal structures characterized as mG receptor of subtype 3, which comprises the sequence with UniprotKB Id P31422 as chain A and B. Table 9.11 shows an overview of the five different crystal structures, which are available for the sequence. A detailed analysis of the n -grams shows that the n -gram SDGW and KIMFV are already known as binding site for Glutamate. N-gram CDAM is involved in a DSB. The other n -grams of table 9.12 are not attributed currently a known functionality.

Structure	Description
2E4U	Extracellular region of group II mG receptor complexed with L-glutamate
2E4V	Extracellular region of group II mG receptor complexed with DCG-IV
2E4W	Extracellular region of group II mG receptor complexed with 1S,3S-ACPD
2E4X	Extracellular region of group II mG receptor complexed with 1S,3R-ACPD
2E4Y	Extracellular region of group II mG receptor complexed with 2R,4R-APDC

Table 9.11.: Crystal structures for mG receptor subtype 3.

N-gram	Pos.	Binding Site	DSB	Structure
RYD	183-185	-	-	-
WTYVS	210-214	-	-	conn. to strand
SDGW	300-303	301, Glutamate	-	-
NHRN	345-348	-	-	-
RNPWF	347-351	-	-	conn. to helix
KIMFV	389-393	389, Glutamate	-	helix
CDAM	419-422	-	412-419	helix
GRY	464-466	-	-	conn. to strand

Table 9.12.: Detailed analysis of functions and structures related to P31422.

Crystal structures of Metabotropic Glutamate Receptor subtype 3 (related to sequence Q14832) In this section we review the crystal structures characterized as mG receptor of subtype 3, which comprises the sequence with UniprotKB Id Q14832 as chain A and B. Table 9.13 shows an overview of the four different crystal structures, which are available for the sequence. A detailed analysis of the *n*-grams shows that the *n*-gram SDGW and KIMFV are already known as binding

site for Glutamate. N-gram CDAM is involved in a DSB. The other n -grams of Table 9.14 are not attributed currently a known functionality.

Structure	Description
3SM9	Metabotropic glutamate receptor 3 precursor in presence of LY341495 antagonist
5CNK	Metabotropic glutamate receptor 3 with Glutamate
5CNM	Metabotropic glutamate receptor 3 complexed with Glutamate analog
4XAR	mGluR2 ECD and mGluR3 ECD complexed with ligands

Table 9.13.: Crystal structures for mG receptor subtype 3 (Q14832).

N-gram	Pos.	Binding Site	DSB	Structure
RYD	183-185	-	-	-
WTYVS	210-214	-	-	conn. to strand
SDGW	300-303	301, Glutamate	-	-
NHRN	345-348	-	-	-
RNPWF	347-351	-	-	conn. to helix
KIMFV	389-393	389, Glutamate	-	helix
CDAM	419-422	-	412-419	helix
GRY	464-466	-	-	conn. to strand

Table 9.14.: Detailed analysis of functions and structures related to Q14832.

Crystal structures of Metabotropic Glutamate Receptor subtype 5 (related to sequence P41594) In this section we review the crystal structures characterized as mG receptor of subtype 5, which comprises the sequence with UniprotKB Id P41594 as chain A and B. Table 9.15 shows the description of the crystal structure available for the sequence. A detailed analysis of the n -grams shows that the n -gram SDGW and KMGFV are already known as binding site for Glutamate.

N-gram CDAM is involved in a DSB. The other n -grams of Table 9.16 are not attributed currently a known functionality.

Structure	Description
3LMK	Ligand Binding Domain of Metabotropic glutamate receptor mGluR5 complexed with glutamate

Table 9.15.: Crystal structures for mG receptor subtype 5 (P41594).

N-gram	Pos.	Binding Site	DSB	Structure
WTYVS	211-215	-	-	conn. to strand
CEGM	278-281			conn. to helix
SDGW	304-307	305, Glutamate	-	-
RYD	310-312	-	-	helix
NHRN	349-352			
RNPWF	351-355	-	-	conn. to helix
KMGFV	396-400	396, Glutamate	-	helix
CDAM	426-429	-	419-426	helix
GRY	464-466		-	conn. to strand

Table 9.16.: Detailed analysis of functions and structures related to P41594.

Crystal structures of Metabotropic Glutamate Receptor subtype 7 (related to sequence P35400) In this section we review the crystal structures characterized as mG receptor of subtype 7, which comprises the sequence with UniprotKB Id P35400 as chain A and B. Table 9.17 shows the description of the crystal structure available for the sequence. A detailed analysis of the n -grams shows that the n -gram SDSW and KVQFV are already known as binding site for Glutamate. N-gram CPEM is involved in a DSB. The other n -grams of Table 9.18 are not attributed currently a known functionality.

Structure	Description
2E4Z	Ligand-binding region of the group III metabotropic glutamate receptor

Table 9.17.: Crystal structures for mG receptor subtype 7 (P35400).

N-gram	Pos.	Binding Site	DSB	Structure
RYD	191-193	-	-	-
WNYVS	218-222	-	-	conn. to strand
SDSW	313-316	314, Glutamate	-	-
NRRN	358-361	-	-	-
RNVWF	360-364	-	-	conn. to helix
WFAE	363-366	-	-	helix
KVQFV	407-411	407, Glutamate	-	helix
CPEM	437-440	-	430-437	-
GRY	475-477	-	-	conn. to strand
RYD	476-478	-	-	conn. to strand

Table 9.18.: Detailed analysis of functions and structures related to P35400.

Crystal structures of Metabotropic Glutamate Receptor subtype 7 (related to sequence Q14831) In this section we review the crystal structures characterized as mG receptor of subtype 7, which comprises the sequence with UniprotKB Id Q14831 as chain A and B. Table 9.19 shows the description of the two crystal structures available for the sequence. A detailed analysis of the n -grams shows that the n -gram SDSW and KVQFV are already known as binding site for Glutamate. N-gram CPEM is involved in a DSB. The other n -grams of Table 9.20 are not attributed currently a known functionality.

Structure	Description
3MQ4	Metabotropic glutamate receptor 7 complexed with LY341495 antagonist
5C5C	Metabotropic glutamate receptor 7

Table 9.19.: Crystal structures for mG receptor subtype 7 (Q14831).

Crystal structures of the extracellular domain of human calcium sensing receptor (related to sequence P41180) In this section we review the crystal structures characterized as a extracellular CS receptor, which comprises the sequence with UniprotKB Id P41180 as chain A and B. Table 9.21 shows an overview of the four different crystal structures, which are available for the sequences. A detailed analysis of the n -grams shows that the n -gram NFRGFRW is already known as an anion binding site. As well n -gram RHLN is related to a Glycosylation site in position 468. The other n -grams of Table 9.22 are not attributed currently a known functionality.

Crystal structures of the extracellular domain of human Gaba B receptor (related to sequence Q9UBS5 and O75899) In this section we review the crystal structures characterized as a extracellular Gaba B receptor, which comprises the sequence with UniprotKB Id Q9UBS5 as chain A and the sequences O75899 as chain B. Table 9.23 shows an overview of the eleven different crystal structures, which are available for the sequences. A detailed analysis of the n -grams shows that for sequence Q9UBBS5 only n -gram DARIIVG is already known as a binding site for agonists and involved in a DSB. For sequence O75899 the n -gram FC-CAY is involved in a DSB. The resting n -grams of Table 9.24 are not attributed currently a known functionality.

9.2.3. Discussion

In the current experiment we have recollected a set of crystal structures related to the extracellular domain of receptors for class C GPCR subtypes mG, CS and GB. For subtype Ta there are no crystal structures available for the extracellular domain.

As a first step we analyzed whether the subtype specific n -grams identified by the χ^2 filtering approach are part of the crystalized sequences. This was thought as a

screening method to identify those crystal structures relevant for a more detailed analysis in reference to the subtype specific n -grams. In the case of mG the analysis was quite successful as 10 out of the 14 crystal structures contained a larger set of subtype specific n -grams. For subtype CS only two crystalized sequences were available, which both contained CS subtype specific n -grams. For GB two out of the five crystalized sequences contained subtype specific n -grams.

A detailed analysis with respect to functional and structural information, either obtained from a protein knowledgebase (UniProtKB) or by visualization of the 3-D structure with UCSF Chimera, revealed interesting information about the subtype specific n -grams: For mG most crystalized sequences contained a n -gram matching the SDXW and KXXFV pattern: For sequences P23385, Q14416, P31422, Q14832, P41594 the first pattern is expressed as n -gram SDGW and for sequences P35400 and Q14831 (mG subtype 7) it is expressed as n -gram SDSW. The second pattern KXXVF for sequences P23385, Q14416, P31422, Q14832 and P41594 appears either as n -gram KMGFV or KIMFV while for sequences P35400 and Q14831 (mG subtype 7) it appears as n -gram KVQFV. This coincidence is important as both patterns describe two known Glutamate binding sites of the mG receptor. As well in most crystalized sequences the n -grams CDAM or CPEM (described with the pattern CXXM) are present, which have a structural importance as part of a DSB. For mG the other subtype specific n -grams have no annotated known functionality. For most mG specific n -grams it was possible to retrieve information about its secondary structure.

For CS two already known binding sites were found: The n -gram NFRGFRW contains a 3-lengths binding site for anions and the n -gram RHLN is a binding site for Glycosylation. The other eight CS specific n -grams do not have a known functionality.

For GB a known agonist binding site was detected with pattern DARIIVG. The other n -grams present in the crystalized structures may also have a functional or structural significance what possibly could be analyzed with the information from the 3-D structure.

9.2.4. Conclusion

In this research we connected the discovery of short subtype specific AA sequences through ML methods with the functional and structural information derived from crystal structure or annotations of a protein knowledge database. In previous research we applied ML methods on the unaligned AA sequences of class C GPCRs and identified for some subtypes, namely mG, CS, GB and Ta, a set of characteristic n -grams. In this investigation we have verified that many of the subtype

specific n -grams actually form part of the crystalized sequences, what makes an analysis regarding their structural and functional significance at the biochemical level feasible. As well we have confirmed the effectiveness of the ML approach used to identify subtype specific n -grams, as some of the detected n -grams describe important segments of the receptors such as binding sites for receptor activation.

A future line of work could be the investigation of those subtype specific n -grams without yet known functionality to determine whether they have a functional or structural significance. This analysis should be done by biochemical experts taking into account the available information about their 3-D configuration from the crystal structures.

Id	Subtype	Sequence	Size	Index	3D structures
M1	mG1	P23385	490	33-522	1EWK, 1EWT, 1EWV 1ISR, 1ISS
M2	mG1	Q13255	496	28-518	3KSG
M3	mG1	Q13255	389	581-860	4OR2
M4	mG2	Q14416	503	2-493	4XAQ, 4XAS, 5CNI 5CNJ
M5	mG2	Q14416	570	1-562	5KZN, 5KZQ
M6	mG3	P31422	555	25-575	2E4U, 2E4V, 2E4W 2E4X, 2E4Y
M7	mG3	Q14832	479	26-504	3SM9
M8	mG3	Q14832	506	2-507	5CNK, 5CNM
M9	mG3	Q14832	507	2-508	4XAR
M10	mG5	P31424	6	1155-1160	1DDV
M11	mG5	P41594	444	569-836	4OO9, 5CGC, 5CGD
M12	mG5	P41594	492	18-505	3LMK
M13	mG7	P35400	501	33-521	2E4Z
M14	mG7	Q14831	481	37-513	3MQ4, 5C5C

Table 9.1.: Relation of crystal structures for mG: Sequence denotes the protein sequence by its Uniprot Accession Id, size denotes the lengths in AA of the segment, Index describes which part of the sequence is crystalized and 3-D structures reports the PDB entry code of the structures.

N-grams										
	RNXWF	GRY	RYD	CXWXC	CXXM	KXXFV	WXYVS	SDXW	NXRN	WFXE
	RNXW			CCWXC			WXXVS			
	NXWF			CCW						
	RXXWF			CWXC						
	RNXXE			WXC						
M1	RNPWF	GRY	RYD (2)		CDAM	KMGFV	WTYVS	SDGW	NTRN	WFPE
M2					CEGM	KSSFV				WFDE
M3										
M4	RNPWF	GRY (2)	RYD		CDAM					
M5	RNPWF	GRY (2)	RYD	CCWLC	CDAM	KIMFV	WTYVS	SDGW		WFRE
M6	RNPWF	GRY	RYD	CCWIC	CDAM	KIMFV	WTYVS	SDGW	NSRN	WFRE
M7	RNPWF	GRY	RYD		CDAM	KIMFV	WTYVS	SDGW	NHRN	
M8	RNPWF	GRY	RYD		CDAM	KIMFV	WTYVS	SDGW	NHRN	
M9	RNPWF	GRY	RYD		CDAM	KIMFV	WTYVS	SDGW	NHRN	
M10						KIMFV	WTYVS	SDGW	NHRN	
M11										
M12	RNPWF	GRY	RYD		CDAM					
					CEGM					
M13	RNVWF	GRY	RYD (2)		CPEM	KMGFV	WTYVS	SDGW	NHRN	
M14	RNVWF	GRY	RYD (2)		CPEM	KVQFV	WNYVS	SDSW	NRRN	WFAE
	RNVNF					KVQFV	WNYVS	SDSW	NRRN	WFAE

Table 9.2.: Analysis of matching betw. mG specific n -grams and crystalized sequences identified with M1- M14.

Id	Subtype	Sequence	Size	Index	3D structures
C1	CS	P41180	568	20- 541	5FBH, 5FBK
C2	CS	P41180	615	20-607	5K5S, 5K5T

Table 9.3.: Relation of crystal structures for CS: Sequence denotes the protein sequence by its Uniprot Accession Id, size denotes the lengths in AA of the segment, Index describes which part of the sequence is crystalized and 3-D structures reports the PDB entry code of the structures.

	N-grams				
	GTRKG	RGFRW	TAXXI	WNWXG	IXXIE
	GTXKG	GFRW	VIVVF	WNXXG	MIXXI
	GTRXG	RGFR	VIVV	NWXG	
	TRKGI	NFRGF	VIXVF	WXGXI	
	TRKXI	FRGFR	VIVXF		
	TXKGI				
C1		NFRGFRW	TAKVIVVF	WNWVGTI	MIFAIE
C2	GTRKGI	NFRGFRW	TAKVIVVF	WNWVGTI	MIFAIE
	N-grams				
	KXIE	GGTIG	AADDD	MAXXI	RXLN
		GGXIG	DDDXG		
C1		GGTIG	AADDDYG	MADII	RHLN
C2	KAIE	GGTIG	AADDDYG	MADII	RHLN

Table 9.4.: Analysis of matching betw. CS specific *n*-grams and crystalized sequences identified with C1 and C2.

Id	Subtype	Sequence	Size	Index	3D structures
G1	GB1	Q9ZOU4	68	96-159	1SRZ, 1SS2
G2	GB1	Q9UBS5	420	165-576	4MQE, 4MQF, 4MR7, 4MR8 4MR9, 4MRM, 4MS1, 4MS3 4MS4
G3	GB1	Q9UBS5	41	879-919	4PAS
G4	GB2	O75899	423	42-466	4F11, 4F12, 4MQE, 4MQF, 4MR7, 4MR8, 4MR9, 4MRM 4MS1, 4MS3, 4MS4
G5	GB2	O75899	41	779-819	4PAS

Table 9.5.: Relation of crystal structures for GB: Sequence denotes the protein sequence by its Uniprot Accession Id, size denotes the lengths in AA of the segment, Index describes which part of the sequence is crystalized and 3-D structures reports the PDB entry code of the structures.

	N-grams				
	YDXXW YDAXW YXAXW DAXW	SKXHG KXHG	DXRII DXRXI RII DXRI RIIXG	GWY	WAXAL WXXAL
G1	YDAIW		DARIIVG	GWY	WALAL
G2					
G3					
G4		SKFHG	DVRIILG	GWY	
	N-grams				
	DGXW YXGXW YDGXW YDXIW YDXXW	FCXXY	WIXXG WIXPG WIIP WIXP YXWII YXWI	AXXVF	
G1	YDGIW				
G2		FCEVY		ARKVF	
G3					
G4		FCCAY	YQWIIIP	AAKVF	

Table 9.6.: Analysis of matching betw. GB specific n -grams and crystalized sequences identified with G1 - G4.

Structure	Description
1EWK	mG receptor subtype 1 complexed with Glutamate
1EWT	mG receptor subtype 1 ligand free form I
1EWW	mG receptor subtype 1 ligand free form II
1ISR1	mG receptor complexed with Glutamate and Gadlinium Ion
1ISS	mG receptor subtype 1 complexed with antagonist

Table 9.7.: Crystal structures for mG receptor subtype 1.

N-gram	Pos.	Binding Site	DSB	Structure
WTYVS	224-228	-	-	from 226 strand
SDGW	317-320	318, Glutamate	-	strand
NTRN	362-365	-	-	-
RNPWF	364-368	-	-	from 367 helix
WFPE	367-370	-	-	helix
KMGFV	409-413	409, Glutamate	-	helix
CDAM	439-442	-	-	strand
KSSFV	456-460	-	-	conn. helix to strand
WFDE	468-471	-	-	strand
GRYD	477-480	-	-	strand

Table 9.8.: Detailed analysis of functions and structures related to P23385.

N-gram	Pos.	Binding Site	DSB	Structure
RYD	191-193	-	-	-
WNYVS	218-222	-	-	conn. to strand
SDSW	313-316	314, Glutamate	-	-
NRRN	358-361	-	-	-
RNVWF	360-364	-	-	conn. to helix
WFAE	363-366	-	-	helix
KVQFV	407-411	407, Glutamate	-	helix
CPEM	437-440	-	430-437	helix
GRY	475-477		-	conn. to strand
RYD	476-478	-	-	conn. to strand

Table 9.20.: Detailed analysis of functions and structures related to Q14831.

Structure	Description
5FBH	Extracellular domain of human caSe receptor with bound Gd3+
5FBK	Extracellular domain of human caSe receptor
5K5S	Active form of human CaSe receptor extracellular domain
5K5T	Inactive form of human CaSe receptor extracellular domain

Table 9.21.: Crystal structures for CS extracellular domain.

N-gram	Pos.	Binding Site	DSB	Structure
NFRGFRW	64-70	66-70, anion binding	-	helix
MIFAIE	75-79	-	-	helix
MADII	197-201	-	-	helix
WNWVGTI	206-212	-	-	partly strand
AADDDYG	213-219	-	-	conn. strand to helix
TAKVIVVF	263-270	-	-	end is strand
GGTIG	315-319	-	-	conn. helix to strand
RHLN	465-468	468, Glycosylation N-linked asparagine	-	conn. helix to strand
GTRKGI	549-554	-	-	strand
KEIE	600-603	-	-	from 602/603 without crystal structure

Table 9.22.: Detailed analysis of functions and structures related to P41180

Structure	Description	Q9UB5S	O75899
4F11	GABA(B) receptor GBR2.		x
4F12	GABA(B) receptor GBR2.		x
4MQE	GABA(B) receptor in the appo form	x	x
4MQF	GABA(B) bound to antagonist 2-hydroxysaclofen	x	x
4MR7	GABA(B) bound to antagonist CGP54626	x	x
4MR8	GABA(B) bound to antagonist CGP35348	x	x
4MR9	GABA(B) bound to antagonist SCH50911	x	x
4MRM	GABA(B) bound to antagonist phaclofen	x	x
4MS1	GABA(B) bound to antagonist CGP46381	x	x
4MS3	GABA(B) bound to endogenous agonist GABA	x	x
4MS4	GABA(B) bound to agonist baclofen	x	x

Table 9.23.: Crystal structures for GB extracellular domain.

Sequence	N-gram	Pos	Binding Site	DSB	Structure
Q9UBS5	DARIIVG	241-247	247, agonist	220-246	helix
	ARKVF	254-258	-	-	helix
	FCEVY	258-262	-	-	helix
	GWY	277-279	-	-	-
	YDAIW	354-358	-	-	helix
	WALAL	358-362	-	-	helix
O75899	DVRIILG	247-253	-		helix to strand
	AAKVF	260-264	-		helix
	FCCAY	264-268	-	265-302	helix
	YQWIIPG	277-283	-		strand
	GWY	283-285	-		-
	SKFHG	352-356	-		partly helix
	YDGIW	359-363	-		helix

Table 9.24.: Detailed analysis of functions and structures related to Q9UBS5

10. Conclusions and future work

We end this dissertation by proposing some lines for future research that arise from the main contributions of the thesis.

In this research we analyzed different types of FS methods in order to examine extensive feature sets of AA patterns. The experiments found a two-step FS approach using forward selection and an univariate test suitable for reducing the feature sets to a small set of very discriminative AA patterns. We applied the proposed approach on both the entire sequence and the N-terminus, which was found to be nearly as discriminative between subtypes as the whole sequence. Focusing on the N-terminus the objective was to find motifs, which form part of the orthosteric binding site in the extracellular domain. The orthosteric binding site is located in the extracellular domain, specifically at the VFT, which comprises two opposing lobes with a cleft where endogenous ligands bind.

As first line of future research we plan to apply the here proposed FS approach also to the segments relevant for the allosteric modulation, i.e. to the 7TM domain. Allosteric modulators are of especial interest in comparison to orthosteric ligands due to their reduced desensitization, tolerance and side effects as well as higher selectivity among receptor subtypes and activity depending on the spatial and temporal presence of endogenous agonist [28]. Furthermore it is worth noting that, although no GPCR allosteric modulators have yet been approved for psychiatric or neurological disorders, a number of GPCR allosteric modulators including, particularly, some from class C, are under clinical development [28].

Regarding the analysis of the significance at the biochemical level of the motifs of the N-terminus we started to analyze whether the short AA patterns identified as subtype specific form part of the set of crystallographic structures known for class C GPCRs. This analysis found some motifs to appear very frequently in the known crystal structures and a short review of the functional annotation of the sequences revealed that many of them do not have a known functionality. The frequent appearance of these AA patterns in the known crystal structures gives foundation to further investigate the significance of these AA patterns in the field of computational chemistry as a second future line of research.

As third line of research we also plan to extend the research on the proposed systematic mislabel analysis approach. In this study, we have proposed a system-

atic procedure, based on SVM classification, to single out and characterize GPCR sequences with consistent misclassification behaviour. The reported experimental results represented a proof of concept for the viability of such procedure as part of a decision support system that, combined with expert knowledge in the field, should be able to assist the discovery of GPCR database labelling quality problems. We plan to extend this research by implementing the proposed mislabel analysis approach as a software tool, what allows an interactive exploration of the mislabeled items for a user defined thresholds for both the *voting ratio* (R_s) and the *cumulative decision value* (CDV_s) .

Bibliography

- [1] V. Katritch, V. Cherezov, and R. C. Stevens, “Structure-function of the G Protein Coupled Receptor Superfamily,” *Annual Review of Pharmacology and Toxicology*, vol. 53, no. 1, pp. 531–556, 2013.
- [2] B. Trzaskowski, D. Latek, S. Yuan, U. Ghoshdastider, A. Debinski, and S. Filipek, “Action of molecular switches in GPCRs-theoretical and experimental studies,” *Current medicinal chemistry*, vol. 19, no. 8, pp. 1090–1109, 2012.
- [3] C. König, R. Cruz-Barbosa, R. Alquézar, and A. Vellido, “SVM-based classification of class C GPCRs from alignment-free physicochemical transformations of their sequences,” in *International Conference on Image Analysis and Processing*, pp. 336–343, Springer, 2013.
- [4] C. König, A. Vellido, R. Alquézar, and J. Giraldo, “Misclassification of class C G-protein coupled receptors as a label noise problem,” in *ESANN 2014: 22st European Symposium on Artificial Neural Networks, Computational Intelligence And Machine Learning: Bruges April 23-24-25, 2014: proceedings*, pp. 695–700, 2014.
- [5] C. König, M. I. Cárdenas, J. Giraldo, R. Alquézar, and A. Vellido, “Label noise in subtype discrimination of class C G protein-coupled receptors: A systematic approach to the analysis of classification errors,” *BMC Bioinformatics*, vol. 16, no. 1, p. 314, 2015.
- [6] M. I. Cárdenas Domínguez, A. Vellido, C. König, R. Alquézar, and J. Giraldo, “Visual characterization of misclassified class C GPCRs through manifold-based machine learning methods,” *Genomics and Computational Biology*, vol. 1, no. 1, p. e19, 2015.
- [7] C. König, I. Shaim, A. Vellido, E. Romero, R. Alquézar, and J. Giraldo, “Using machine learning tools for protein database biocuration assistance,” *Scientific reports*, vol. 8, no. 1, p. 10148, 2018.
- [8] C. König, R. Alquézar, A. Vellido, and J. Giraldo, “Topological Sequence Segments Discriminate Between Class C GPCR Subtypes,” in *International Conference on Practical Applications of Computational Biology & Bioinformatics*, pp. 164–172, Springer, 2017.

- [9] C. König, R. Alquézar, A. Vellido, and J. Giraldo, “Systematic Analysis of Primary Sequence Domain Segments for the Discrimination Between Class C GPCR Subtypes,” *Interdisciplinary Sciences: Computational Life Sciences*, vol. 10, no. 1, pp. 43–52, 2018.
- [10] C. König, R. Alquézar, A. Vellido, and J. Giraldo, “Finding Class C GPCR Subtype-Discriminating N-grams through Feature Selection,” in *8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014)*, pp. 89–96, Springer, 2014.
- [11] C. König, R. Alquézar, A. Vellido, and J. Giraldo, “Reducing the n-gram feature space of class C GPCRs to subtype-discriminating patterns,” *J. Integrative Bioinformatics*, vol. 11, no. 3, 2014.
- [12] C. König, R. Alquézar, A. Vellido, and J. Giraldo, “The extracellular N-terminal domain suffices to discriminate class C G Protein-Coupled Receptor subtypes from n-grams of their sequences,” in *2015 International Joint Conference on Neural Networks, IJCNN 2015, Killarney, Ireland, July 12-17, 2015*, pp. 1–7, 2015.
- [13] C. König, R. Alquézar, A. Vellido, and J. Giraldo, “Discovering Subtype Specific n-Gram Motifs in Class C GPCR N-Termini,” in *Recent Advances in Artificial Intelligence Research and Development - Proceedings of the 20th International Conference of the Catalan Association for Artificial Intelligence, Deltebre, Terres de l’Ebre, Spain, October 25-27, 2017*, pp. 116–125, 2017.
- [14] A. Schmidt, I. Forne, and A. Imhof, “Bioinformatic analysis of proteomics data,” *BMC systems biology*, vol. 8, no. 2, p. S3, 2014.
- [15] P. G. Ferreira and P. J. Azevedo, “Evaluating deterministic motif significance measures in protein databases,” *Algorithms for Molecular Biology*, vol. 2, no. 1, p. 16, 2007.
- [16] T. Mi, J. C. Merlin, S. Deverasetty, M. R. Gryk, T. J. Bill, A. W. Brooks, L. Y. Lee, V. Rathnayake, C. A. Ross, D. P. Sargeant, *et al.*, “Minimotif miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences,” *Nucleic acids research*, vol. 40, no. D1, pp. D252–D260, 2011.
- [17] H. Dinkel, K. Van Roey, S. Michael, M. Kumar, B. Uyar, B. Altenberg, V. Milchevskaya, M. Schneider, H. Kühn, A. Behrendt, *et al.*, “Elm 2016 data update and new functionality of the eukaryotic linear motif resource,” *Nucleic acids research*, vol. 44, no. D1, pp. D294–D300, 2015.
- [18] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble, “The meme suite,” *Nucleic acids research*, vol. 43, no. W1, pp. W39–W49, 2015.

- [19] D. La and D. R. Livesay, “Miner: software for phylogenetic motif identification,” *Nucleic acids research*, vol. 33, no. suppl_2, pp. W267–W270, 2005.
- [20] M. Rask-Andersen, M. Sällman Almén, and H. B. Schiöth, “Trends in the exploitation of novel drug targets,” vol. 10, pp. 579–90, 08 2011.
- [21] R. Santos, O. Ursu, A. Gaulton, A. Patrícia Bento, R. S. Donadi, C. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T. Oprea, and J. Overington, “A comprehensive map of molecular drug targets,” vol. 16, no. 1, p. 19, 2017.
- [22] G. Ladds, A. Goddard, and J. Davey, “Functional analysis of heterologous gpcr signalling pathways in yeast,” *Trends in biotechnology*, vol. 23, no. 7, pp. 367–373, 2005.
- [23] A. de Mendoza, A. Sebé-Pedrós, and I. Ruiz-Trillo, “The Evolution of the GPCR Signaling System in Eukaryotes: Modularity, Conservation, and the Transition to Metazoan Multicellularity,” vol. 6, no. 3, pp. 606–619, 2014.
- [24] K. Leach and K. J. Gregory, “Molecular insights into allosteric modulation of Class C G protein-coupled receptors,” *Pharmacological Research*, vol. 116, pp. 105 – 118, 2017.
- [25] J. Kniazeff, L. Prézeau, P. Rondard, J.-P. Pin, and C. Goudet, “Dimers and beyond: The functional puzzles of class C GPCRs,” *Pharmacology Therapeutics*, vol. 130, no. 1, pp. 9 – 25, 2011.
- [26] J.-P. Pin, T. Galvez, and L. Prézeau, “Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors,” *Pharmacology & Therapeutics*, vol. 98, no. 3, pp. 325 – 354, 2003.
- [27] J. Cao, S. Huang, J. Qian, J. Huang, L. Jin, Z. Su, J. Yang, and J. Liu, “Evolution of the class C GPCR Venus flytrap modules involved positive selected functional divergence,” *BMC Evolutionary Biology*, vol. 9, no. 1, p. 67, 2009.
- [28] P. Conn, C. Lindsley, J. Meiler, and C. M. Niswender, “Opportunities and Challenges in the Discovery of Allosteric Modulators of GPCRs for Treating CNS Disorders,” vol. 13, pp. 692–708, 09 2014.
- [29] V. Isberg, B. Vroling, R. van der Kant, K. Li, G. Vriend, and D. Gloriam, “GPCRDB: an information system for G protein-coupled receptors,” *Nucleic acids research*, vol. 42, no. D1, pp. D422–D425, 2014.
- [30] F. Nicoletti, J. Bockaert, G. Collingridge, P. Conn, F. Ferraguti, D. Schoepp, J. Wroblewski, and J. Pin, “Metabotropic glutamate receptors: From the workbench to the bedside,” *Neuropharmacology*, vol. 60, no. 7, pp. 1017 – 1041, 2011.

- [31] F. Nicoletti, V. Bruno, R. T. Ngomba, R. Gradini, and G. Battaglia, “Metabotropic glutamate receptors as drug targets: what’s new?,” *Current Opinion in Pharmacology*, vol. 20, no. Supplement C, pp. 89 – 94, 2015.
- [32] K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. L. Trong, D. C. Teller, T. Okada, R. E. Stenkamp, M. Yamamoto, and M. Miyano, “Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor,” *Science*, vol. 289, no. 5480, pp. 739–745, 2000.
- [33] V. Katritch, V. Cherezov, and R. C. Stevens, “Structure Function of the G Protein Coupled Receptor Superfamily,” *Annual Review of Pharmacology and Toxicology*, vol. 53, no. 1, pp. 531–556, 2013.
- [34] S. Alexander, A. P Davenport, E. Kelly, N. Marrion, J. Peters, H. Benson, E. Faccenda, A. J Pawson, J. Sharman, C. Southan, and J. Davies, “The concise guide to PHARMACOLOGY 2015/16: G protein-coupled receptors,” vol. 172, pp. 5744–5869, 2015.
- [35] R. Cooke, A. Brown, F. Marshall, and J. Mason, “Structures of G-protein-coupled receptors reveal new opportunities for drug discovery,” vol. 20, no. 11, pp. 1355–1364, 2015.
- [36] J. L. Sharman and C. P. Mpamhanga, “IUPHAR-DB: an open-access, expert-curated resource for receptor and ion channel research,” *ACS Chemical Neuroscience*, vol. 2, pp. 232–235, 2011.
- [37] C. Munk, V. Isberg, S. Mordalski, K. Harpsøe, K. Rataj, A. Hauser, P. Kolb, A. Bojarski, G. Vriend, and D. Gloriam, “GPCRdb: the G protein-coupled receptor database - an introduction,” *British Journal of Pharmacology*, vol. 173, no. 14, pp. 2195–2207, 2016.
- [38] V. Isberg, S. Mordalski, C. Munk, K. Rataj, K. Harpsøe, A. S. Hauser, B. Vroiling, A. J. Bojarski, G. Vriend, and D. E. Gloriam, “GPCRdb: an information system for G protein-coupled receptors,” *Nucleic Acids Research*, vol. 45, no. 5, p. 2936, 2017.
- [39] R. Fredriksson, M. C. Lagerström, L.-G. Lundin, and H. B. Schiöth, “The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints,” *Molecular Pharmacology*, vol. 63, no. 6, pp. 1256–1272, 2003.
- [40] L. F. Kolakowski Jr, “GCRDb: a G-protein-coupled receptor database.,” *Receptors & Channels*, vol. 2, no. 1, pp. 1–7, 1993.
- [41] J.-P. Pin and B. Bettler, “Organization and functions of mglu and GABAB receptor complexes,” vol. 540, pp. 60–68, 12 2016.

- [42] R. Santamaría and R. Therón, “Treevolution: visual analysis of phylogenetic trees,” *Bioinformatics*, vol. 25, no. 15, pp. 1970–1971, 2009.
- [43] A. R. Shah, C. S. Oehmen, and B.-J. Webb-Robertson, “Svm HUSTLE an iterative semi-supervised machine learning approach for pairwise protein remote homology detection,” *Bioinformatics*, vol. 24, no. 6, p. 783, 2008.
- [44] Y. Hou, W. Hsu, M. L. Lee, and C. Bystroff, “Efficient remote homology detection using local structure,” *Bioinformatics*, vol. 19, no. 17, p. 2294, 2003.
- [45] H. Ogul and E. Mumcuoglu, “A discriminative method for remote homology detection based on n-peptide compositions with reduced amino acid alphabets,” *Biosystems*, vol. 87, no. 1, pp. 75 – 81, 2007.
- [46] B.-J. Webb-Robertson, C. Oehmen, and M. Matzke, “Svm-balsa: Remote homology detection based on bayesian sequence alignment,” *Computational Biology and Chemistry*, vol. 29, no. 6, pp. 440 – 443, 2005.
- [47] R. Karchin, K. Karplus, and D. Haussler, “Classifying G-protein coupled receptors with support vector machines,” *Bioinformatics*, vol. 18, no. 1, p. 147, 2002.
- [48] B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan, “Using amino acid physicochemical distance transformation for fast protein remote homology detection,” *PLOS ONE*, vol. 7, pp. 1–10, 09 2012.
- [49] S. O. Opiyo and E. N. Moriyama, “Protein family classification with partial least squares,” *Journal of Proteome Research*, vol. 6, no. 2, pp. 846–853, 2007.
- [50] B. Y. M. Cheng, J. G. Carbonell, and J. Klein-Seetharaman, “Protein classification based on text document classification techniques,” *Proteins: Structure, Function, and Bioinformatics*, vol. 58, no. 4, pp. 955–970, 2005.
- [51] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.
- [52] J. R. Quinlan, “C4.5: Programs for Machine Learning,” *Machine Learning*, vol. 16, no. 3, pp. 235–240, 1993.
- [53] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [54] T. G. Dietterich, “Ensemble methods in machine learning,” in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.
- [55] G. H. John and P. Langley, “Estimating Continuous Distributions in Bayesian Classifiers,” in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI’95, pp. 338–345, 1995.

- [56] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [57] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A Training Algorithm for Optimal Margin Classifiers,” in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pp. 144–152, ACM, 1992.
- [58] M. A. Aizerman, E. M. Braverman, and L. I. Rozoner, “Theoretical foundations of the potential function method in pattern recognition learning,” *Automation and Remote Control*, vol. 25, pp. 821–837, 1964.
- [59] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
- [60] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, “Evaluation measures for models assessment over imbalanced data sets,” *Journal Of Information Engineering and Applications*, vol. 3, no. 10, 2013.
- [61] G. M. Weiss and F. Provost, “Learning when Training Data Are Costly: The Effect of Class Distribution on Tree Induction,” *Journal of Artificial Intelligence Research*, vol. 19, no. 1, pp. 315–354, 2003.
- [62] M. Sokolova and G. Lapalme, “A Systematic Analysis of Performance Measures for Classification Tasks,” *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, 2009.
- [63] G. Jurman, S. Riccadonna, and C. Furlanello, “A comparison of MCC and CEN error measures in multi-class prediction,” *PloS one*, vol. 7, no. 8, p. e41882, 2012.
- [64] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001.
- [65] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, “Mismatch string kernels for discriminative protein classification,” *Bioinformatics*, vol. 20, no. 4, pp. 467–476, 2004.
- [66] C. Leslie, E. Eskin, and W. S. Noble, “The spectrum kernel: a string kernel for svm protein classification,” in *Pacific Symposium on Biocomputing*, vol. 7, pp. 566–575, 2002.
- [67] T. Jaakkola and D. Haussler, “Exploiting generative models in discriminative classifiers,” in *Advances in neural information processing systems*, pp. 487–493, 1999.
- [68] S. Wold, J. Jonsson, M. Sjöström, M. Sandberg, and S. Rännar, “DNA and peptide sequences and chemical processes multivariately modelled by

- principal component analysis and partial least-squares projections to latent structures,” *Analytica Chimica Acta*, vol. 277, no. 2, pp. 239 – 253, 1993.
- [69] M. Lapinsh, A. Gutcaits, P. Prusis, C. Post, T. Lundstedt, and J. E. Wikberg, “Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences.,” *Protein Science*, vol. 11, no. 4, pp. 795–805, 2002.
- [70] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, and S. Wold, “New chemical descriptors relevant for the design of biologically active peptides. a multivariate characterization of 87 amino acids,” *Journal of Medicinal Chemistry*, vol. 41, no. 14, pp. 2481–2491, 1998.
- [71] R. Cruz Barbosa, A. Vellido Alcacena, and J. Giraldo, “Advances in semi-supervised alignment-free classification of G protein-coupled receptors,” in *Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering 2013*, pp. 759–766, 2013.
- [72] R. Cruz-Barbosa, A. Vellido, and J. Giraldo, “The influence of alignment-free sequence representations on the semi-supervised classification of class C G protein-coupled receptors,” *Medical & Biological Engineering & Computing*, vol. 53, no. 2, pp. 137–149, 2015.
- [73] J. S. Bernardes, A. Carbone, and G. Zaverucha, “A discriminative method for family-based protein remote homology detection that combines inductive logic programming and propositional models,” *BMC Bioinformatics*, vol. 12, p. 83, Mar 2011.
- [74] C. Caragea, A. Silvescu, and P. Mitra, “Protein sequence classification using feature hashing,” *Proteome Science*, vol. 10, no. 1, p. S14, 2012.
- [75] J. Cao and L. Xiong, “Protein Sequence Classification with Improved Extreme Learning Machine Algorithms,” in *BioMed research international*, vol. 2014, 2014.
- [76] F. Mhamdi, M. Elloumi, and R. Rakotomalala, “Textmining, feature selection and datamining for proteins classification,” in *Proceedings. 2004 International Conference on Information and Communication Technologies: From Theory to Applications*, pp. 457–458, 04 2004.
- [77] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.
- [78] E. Asgari and M. R. Mofrad, “Continuous distributed representation of biological sequences for deep proteomics and genomics,” *PloS One*, vol. 10, no. 11, p. e0141287, 2015.

- [79] I. Terpugova, “Protein classification from primary structures in the context of database biocuration,” Master’s thesis, Universitat Politècnica de Catalunya, 2017.
- [80] M. N. Davies, A. Secker, A. A. Freitas, E. Clark, J. Timmis, and D. R. Flower, “Optimizing amino acid groupings for GPCR classification,” *Bioinformatics*, vol. 24, no. 18, pp. 1980–1986, 2008.
- [81] M. C. Cobanoglu, Y. Saygin, and U. Sezerman, “Classification of GPCRs using family specific motifs,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 6, pp. 1495–1508, 2011.
- [82] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
- [83] D. Koller and M. Sahami, “Toward optimal feature selection,” tech. rep., Stanford InfoLab, 1996.
- [84] H. Liu, J. Li, and L. Wong, “A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns,” *Genome informatics*, vol. 13, pp. 51–60, 2002.
- [85] J. Kittler, “Feature set search algorithms,” *Pattern recognition and signal processing*, 1978.
- [86] D. B. Skalak, “Prototype and feature selection by sampling and random mutation hill climbing algorithms,” in *Proceedings of the eleventh international conference on machine learning*, pp. 293–301, 1994.
- [87] J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.
- [88] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.
- [89] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [90] T. O. Conrad, M. Genzel, N. Cvetkovic, N. Wulkow, A. Leichtle, J. Vybiral, G. Kutyniok, and C. Schütte, “Sparse proteomics analysis—a compressed sensing-based approach for feature selection and classification of high-dimensional proteomics mass spectrometry data,” *BMC bioinformatics*, vol. 18, no. 1, p. 160, 2017.
- [91] Y. Wang, L. Feng, and Y. Li, “Two-step based feature selection method for filtering redundant information,” *Journal of Intelligent & Fuzzy Systems*, vol. 33, no. 4, pp. 2059–2073, 2017.

- [92] J. D. Storey, “A direct approach to false discovery rates,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 479–498, 2002.
- [93] J. D. Storey and R. Tibshirani, “Statistical significance for genomewide studies,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 16, pp. 9440–9445, 2003.
- [94] B. Frenay and M. Verleysen, “Classification in the Presence of Label Noise: A Survey,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, 2014.
- [95] P. J. Lisboa, A. Vellido, and J. D. Martín-Guerrero, “Computational intelligence in biomedicine: Some contributions.,” in *Procs. of the 18th European Symposium on Artificial Neural Networks (ESANN)*, pp. 429–438, 2010.
- [96] B. Frénay, G. de Lannoy, and M. Verleysen, “Label noise-tolerant hidden Markov models for segmentation: application to ECGs,” *Machine learning and knowledge discovery in databases*, pp. 455–470, 2011.
- [97] Q.-B. Gao, X.-F. Ye, and J. He, “Classifying G-Protein-Coupled Receptors to the finest subtype level,” *Biochemical and biophysical research communications*, vol. 439, no. 2, pp. 303–308, 2013.
- [98] P. A. Nuin, Z. Wang, and E. R. Tillier, “The accuracy of several multiple sequence alignment programs for proteins,” *BMC bioinformatics*, vol. 7, no. 1, p. 471, 2006.
- [99] K. Ye, W. A. Kosters, and A. P. IJzerman, “An efficient, versatile and scalable pattern growth approach to mine frequent patterns in unaligned protein sequences,” *Bioinformatics*, vol. 23, no. 6, pp. 687–693, 2007.
- [100] C. E. Brodley and M. A. Friedl, “Identifying mislabeled training data,” *Journal of artificial intelligence research*, vol. 11, pp. 131–167, 1999.
- [101] S. Verbaeten and A. Van Assche, “Ensemble methods for noise elimination in classification problems,” in *International Workshop on Multiple Classifier Systems*, pp. 317–325, Springer, 2003.
- [102] B. Sluban, D. Gamberger, and N. Lavrač, “Ensemble-based noise detection: noise ranking and visual performance evaluation,” *Data mining and knowledge discovery*, vol. 28, no. 2, pp. 265–303, 2014.
- [103] D. Gamberger and N. Lavrač, “Conditions for Occam’s razor applicability and noise elimination,” *Machine Learning: ECML-97*, pp. 108–123, 1997.
- [104] D. Gamberger, N. Lavrac, and C. Groselj, “Experiments with noise filtering in a medical domain,” in *ICML*, pp. 143–151, 1999.

- [105] G. Libralon, A. Carvalho, and A. Lorena, “Ensembles of pre-processing techniques for noise detection in gene expression data,” *Advances in Neuro-Information Processing*, pp. 486–493, 2009.
- [106] T. M. Khoshgoftaar, S. Zhong, and V. Joshi, “Enhancing software quality estimation using ensemble-classifier based noise filtering,” *Intelligent Data Analysis*, vol. 9, no. 1, pp. 3–27, 2005.
- [107] A. Vellido, E. Romero, F. F. González-Navarro, L. A. Belanche-Muñoz, M. Julià-Sapé, and C. Arús, “Outlier exploration and diagnostic classification of a multi-centre 1 H-MRS brain tumour database,” *Neurocomputing*, vol. 72, no. 13, pp. 3085–3097, 2009.
- [108] S. Dawood, S. Merajver, P. Viens, P. Vermeulen, S. Swain, T. Buchholz, L. Dirix, P. Levine, A. Lucci, S. Krishnamurthy, *et al.*, “International expert panel on inflammatory breast cancer: consensus statement for standardized diagnosis and treatment,” *Annals of Oncology*, vol. 22, no. 3, pp. 515–523, 2010.
- [109] M. E. Cusick, H. Yu, A. Smolyar, K. Venkatesan, A.-R. Carvunis, N. Simonis, J.-F. Rual, H. Borick, P. Braun, M. Dreze, *et al.*, “Literature-curated protein interaction datasets,” *Nature methods*, vol. 6, no. 1, pp. 39–46, 2009.
- [110] B. Vroling, M. Sanders, C. Baakman, A. Borrmann, S. Verhoeven, J. Klomp, L. Oliveira, J. de Vlieg, and G. Vriend, “GPCRDB: information system for G protein-coupled receptors,” *Nucleic Acids Research*, vol. 39, no. Database issue, p. D309, 2011.
- [111] B. Rehm, “Bioinformatic tools for DNA/protein sequence analysis, functional assignment of genes and protein classification,” *Applied microbiology and biotechnology*, vol. 57, no. 5-6, pp. 579–592, 2001.
- [112] J. G. Martinez, R. J. Carroll, S. Müller, J. N. Sampson, and N. Chatterjee, “Empirical performance of cross-validation with oracle methods in a genomics context,” *The American Statistician*, vol. 65, no. 4, pp. 223–228, 2011.
- [113] K. Jayawardana, S.-J. Schramm, L. Haydu, J. F. Thompson, R. A. Scolyer, G. J. Mann, S. Müller, and J. Y. H. Yang, “Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation, mRNA, microRNA, and protein information,” *International Journal of Cancer*, vol. 136, no. 4, pp. 863–874, 2015.
- [114] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, *et al.*, “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega,” *Molecular systems biology*, vol. 7, no. 1, p. 539, 2011.

- [115] D.-F. Feng and R. F. Doolittle, “Progressive sequence alignment as a prerequisite to correct phylogenetic trees,” *Journal of molecular evolution*, vol. 25, no. 4, pp. 351–360, 1987.
- [116] D. F. Nettleton, A. Orriols-Puig, and A. Fornells, “A study of the effect of different types of noise on the precision of supervised learning techniques,” *Artificial intelligence review*, vol. 33, no. 4, pp. 275–306, 2010.
- [117] M. Pechenizkiy, A. Tsymbal, S. Puuronen, and O. Pechenizkiy, “Class noise and supervised learning in medical domains: The effect of feature extraction,” in *Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on*, pp. 708–713, 2006.
- [118] A. Miranda, L. Garcia, A. Carvalho, and A. Lorena, “Use of classification algorithms in noise detection and elimination,” *Hybrid Artificial Intelligence Systems*, pp. 417–424, 2009.
- [119] B. Sluban, D. Gamberger, and N. Lavra, “Advances in class noise detection,” in *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pp. 1105–1106, IOS Press, 2010.
- [120] J. A. Sáez, M. Galar, J. Luengo, and F. Herrera, “Analyzing the presence of noise in multi-class problems: alleviating its influence with the one-vs-one decomposition,” *Knowledge and information systems*, vol. 38, no. 1, pp. 179–206, 2014.
- [121] A. Shkurin and A. Vellido, “Using Random Forests for assistance in the curation of G-protein coupled receptor databases,” *Biomedical Engineering Online*, vol. 16, p. 75, 2017.
- [122] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. Hill, R. Kania, M. Schaeffer, S. St Pierre, and S. Twigger, “Big data: The future of biocuration,” *Nature*, vol. 455, no. 7209, pp. 47–50, 2008.
- [123] A. Baxevanis and A. Bateman, “The importance of biological databases in biological discovery,” *Current protocols in bioinformatics*, vol. 50, pp. 1–1, 2015.
- [124] A. Singhal, R. Leaman, N. Catlett, T. Lemberger, J. McEntyre, S. Polson, I. Xenarios, C. Arighi, and Z. Lu, “Pressing needs of biomedical text mining in biocuration and beyond: opportunities and challenges,” *Database (Oxford)*, vol. 2016, 2016.
- [125] P. Lord, A. Macdonald, L. Lyon, and G. D., “From data deluge to data curation,” in *Proceedings of the UK e-Science All Hands Meeting*, pp. 371–375, 2004.

- [126] F.-R. Meng, Z.-H. You, X. Chen, Y. Zhou, and J.-Y. An, “Prediction of drug target interaction networks from the integration of protein sequences and drug chemical structures,” *Molecules*, vol. 22, p. 1119, 7 2017.
- [127] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak, “Improved prediction of signal peptides: Signalp 3.0,” *Journal of Molecular Biology*, vol. 340, no. 4, pp. 783 – 795, 2004.
- [128] X. Rovira, F. Malhaire, P. Scholler, J. Rodrigo, P. Gonzalez-Bulnes, A. Llebaria, J.-P. Pin, J. Giraldo, and C. Goudet, “Overlapping binding sites drive allosteric agonism and positive cooperativity in type 4 metabotropic glutamate receptors,” *The FASEB Journal*, vol. 29, no. 1, pp. 116–130, 2015.
- [129] R. K. Hart, A. K. Royyuru, G. Stolovitzky, and A. Califano, “Systematic and fully automated identification of protein sequence patterns,” *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 585–600, 2000.
- [130] K. Jain, “Role of pharmacoproteomics in the development of personalized medicine,” *Pharmacogenomics*, vol. 5, no. 3, pp. 331–336, 2004.

A. Appendix

A.1. Figures of the misclassification analysis

This section contains the figures related to the misclassification analysis of the 2011 class C GPCR dataset described in section 5.3.2:

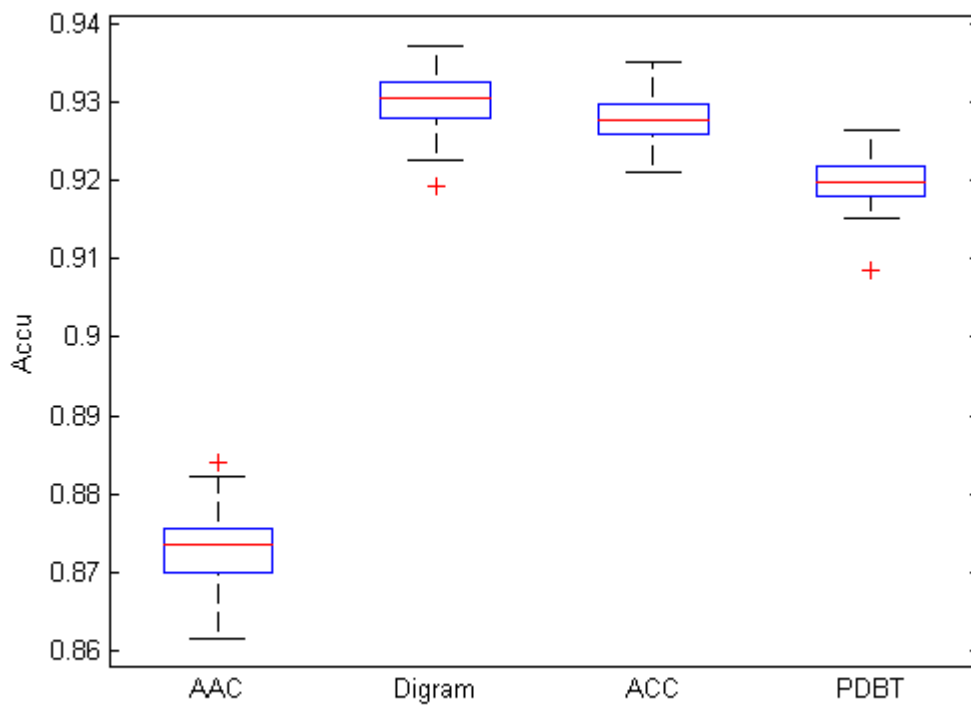


Figure A.1.1.: Boxplot representation of the Accu of the AA, Digram, ACC and PDBT dataset.

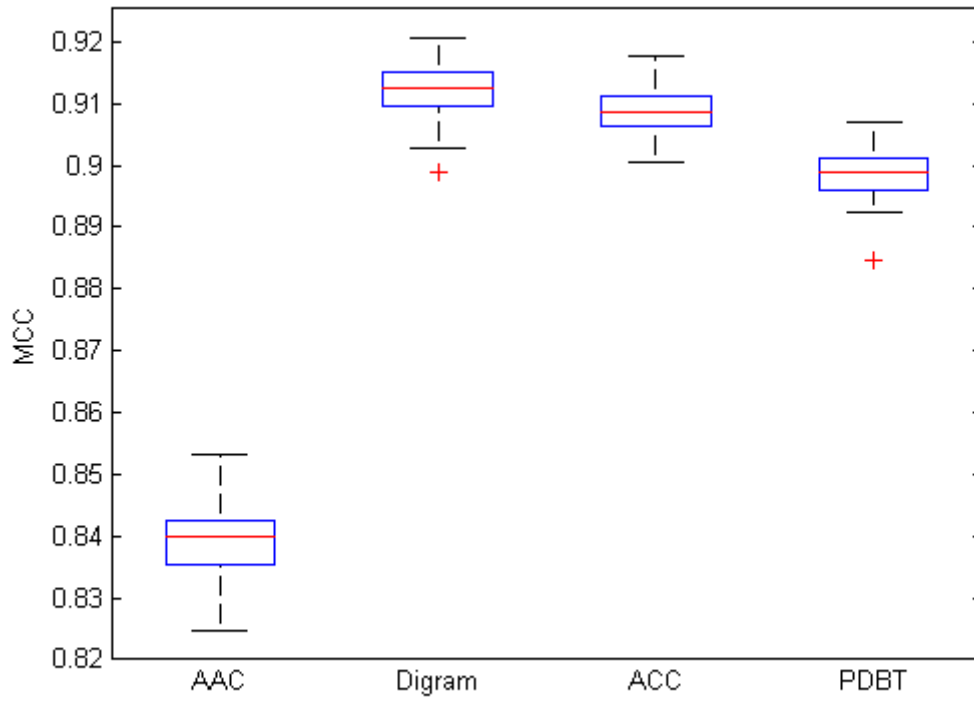


Figure A.1.2.: Boxplot representation of the MCC of the AAC, Digram, ACC and PDBT dataset.

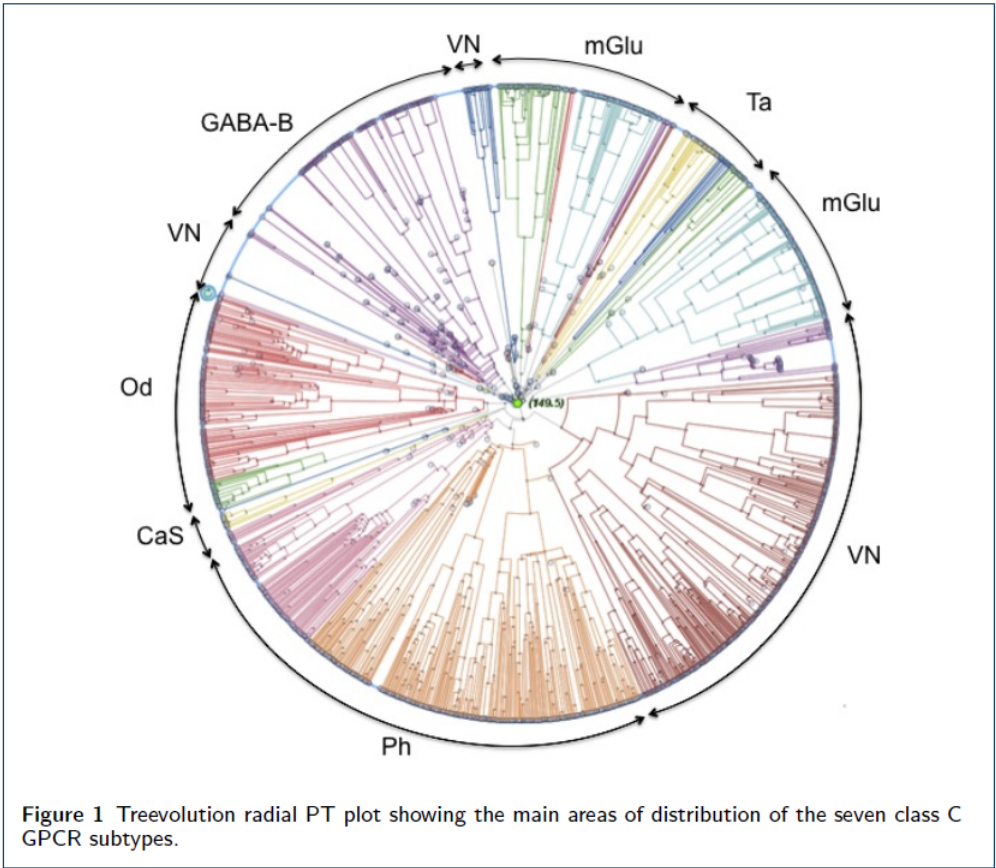


Figure A.1.3.: Treevolution radial PT in which the main sections occupied by each of the seven class C GPCR subtypes are explicitly represented by archs or groups of archs in the periphery of the tree. Note that branch colors are automatically generated during PT construction and do not correspond to class C subtypes.

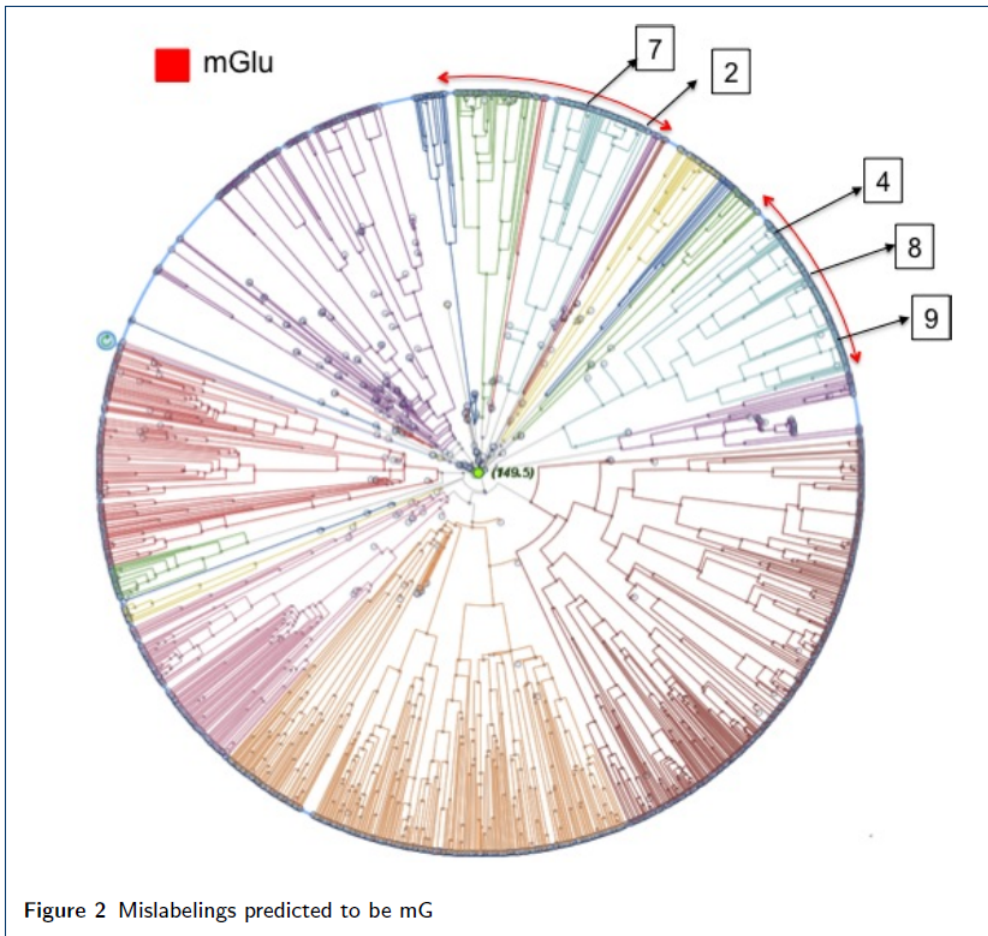


Figure A.1.4.: Mislabeledings predicted to be mG. Five sequences with large classification errors were mislabeled as mG. Sequence #7 was labeled as *VN* in GPCRDB; #2 and #4 were labeled as *CS*; #8 was labeled as *Ph*; and #9 was labeled as *Od*.

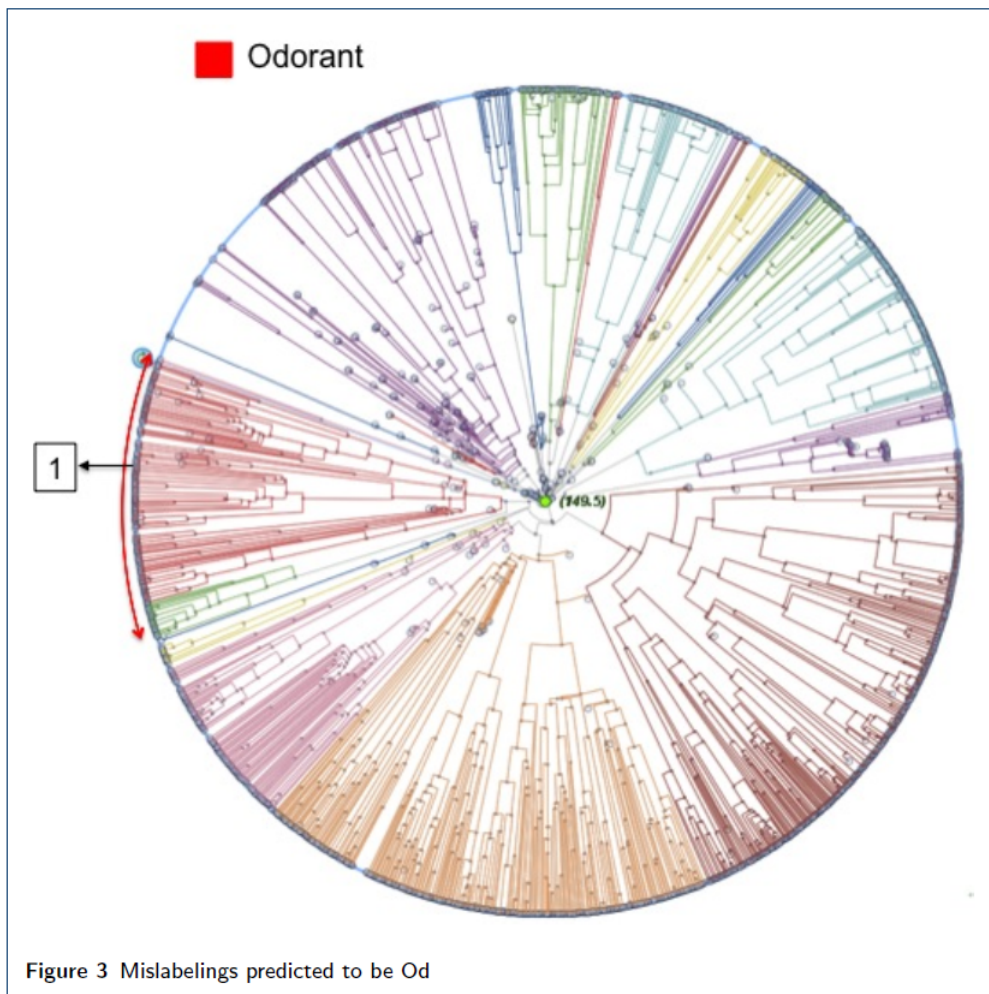


Figure A.1.5.: Mislabelings predicted to be Od. One sequence (#1, labeled as *mG* in GPCRDB) with large classification error was mislabeled as *Od*.

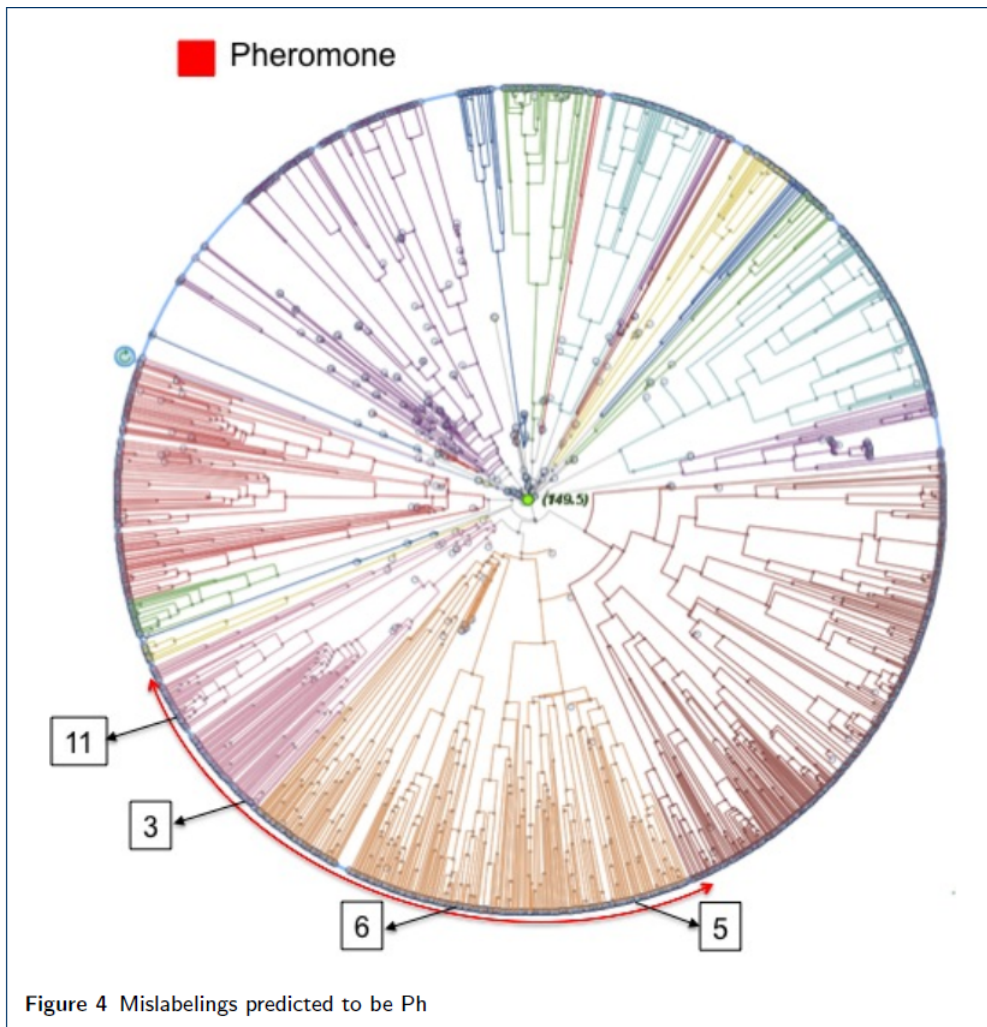


Figure A.1.6.: Mislabeledings predicted to be Ph. Four sequences with large classification errors were mislabeled as Ph. Sequence #3 was labeled as *CS* in GPCRDB; #11 was labeled as *Od*; and #5 and #6 were labeled as *VN*.

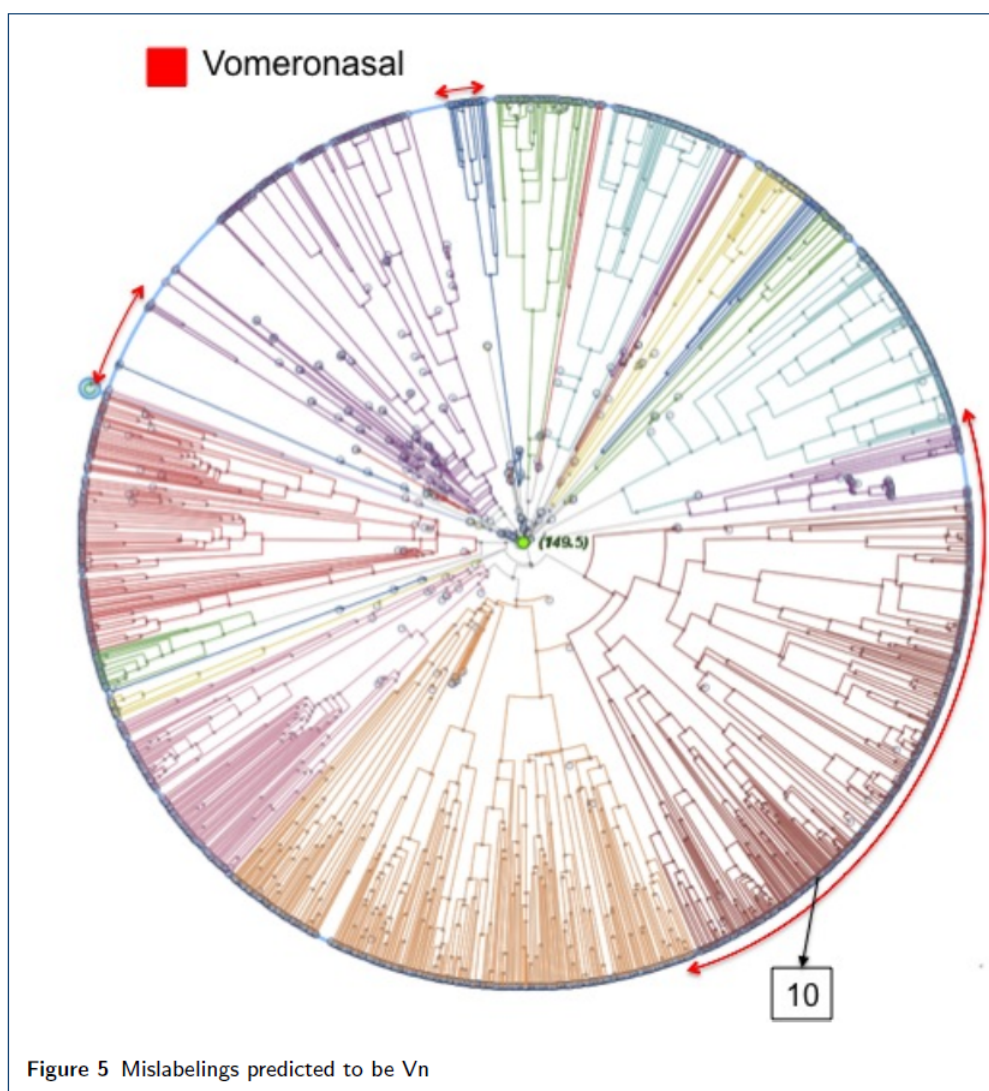


Figure A.1.7.: Mislabelings predicted to be Vn. One sequence (#10, labeled as *Od* in GPCRDB) with large classification error was mislabeled as *Vn*.

A.2. List of frequent misclassified sequences

Tables A.1 and A.2 contain the list of 52 frequently misclassified sequences that were common to all four data transformations, namely the AAC, Digram, ACC and PDBT transformations. A strong agreement on the most-often predicted class C GPCR subtypes can be observed.

	True class	Predicted class label			
GPCRdb Id	GPCRdb	ACC	Digram	ACC	PDBT
a8dz71_danre	mG	Od	Od	Od	Od
a8dz72_danre	mG	Od	Od	Ph	Od
q5i5d4_9tele	mG	Od	Od	Od	Od
q5i5c3_9tele	mG	Od	Od	Od	Od
XP_002123664	CS	mG	mG	mG	mG
q8c0m6_mouse	CS	GB	GB	Ph	GB
XP_002740613	CS	Ph	Od	Od	Ph
XP_002738008	GB	mG	mG	mG	mG
q8bid7_mouse	VN	Od	Od	Od	Od
XP_002936197	VN	Ph	Ph	Ph	Ph
XP_002940476	VN	Ph	Ph	Ph	Ph
XP_0029341318	VN	mG	mG	mG or Ph	mG
XP_002941777	VN	mG	Ph	mG or Ph	Ph
NP_001093066	VN	mG	Ph	Ph	mG
NP_001093039	VN	Ph	Ph	Ph	Ph
XP_001517645	VN	Ph or Od	Ph	Ph	Od
NP_001098007	VN	Ph	Ph	Ph	Ph
o70411_rat	VN	Od	Od	Od	Od
q8tdu1_human	VN	Od	Od	Od	Od
XP_917917	VN	Od	Od	Od	Od
a7sdg9_nemve	Ph	mG	mG	mG	mG
a7s0d2_nemve	Ph	mG	mG	mG	mG
b3s157_triad	Ph	GB	mG	mG	mG
b3s609_triad	Ph	mG	mG	mG	mG
XP_002732067	Ph	mG	mG	mG	mG

Table A.1.: Frequently misclassified sequences common to all four data transformations for subtypes mG, Cs, GB, VN and Ph. It includes the following columns: GPCRDB identifier, GPCRDB true class, and predicted class for, in turn, the AAC, Digram, ACC and PDBT transformations.

	True class	Predicted class label			
GPCRdb Id	GPCRdb	ACC	Digram	ACC	PDBT
q4spr3_tetng	Ph	GB	mG	mG	mG
XP_002937659	Ph	mG	mG	mG	mG
XP_001368172	Ph	VN	VN	VN	VN
NP_001093018	Ph	VN	VN	VN	VN
NP_001093020	Ph	VN	VN	VN	VN
NP_001093016	Ph	VN	VN	VN	VN
XP_002723938	Ph	VN	VN	VN	VN
XP_002936172	Ph	VN	VN	VN	VN
q9pwe1_ictpu	Ph	mG	mG	mG	mG
b0uyj3_danre	Ph	mG	mG	mG	mG
NP_001093040	Ph	VN	VN	VN	VN
XP_001075542	Ph	Od	Od	Od	Od
XP_001521075	Ph	mG	mG	mG	mG
q6unx3_ictpu	Od	Ph	Ph	Ph	mG or Ph
gpc6a_human	Od	VN	Ph	Ph	mG or VN
b3rud8_triad	Od	mG	mG	mG	mG
d1lwx7_sacko	Od	VN	CS	CS	mG or CS
XP_002936177	Od	VN	VN	VN	VN or Ph
XP_002936183	Od	mG or VN	VN	Ph	Ph
XP_002937663	Od	Ph	Ph	Ph	Ph
XP_002940477	Od	Ph	Ph	Ph	Ph
XP_002940566	Od	VN	Ph	Ph	Ph
XP_002940324	Od	VN	VN	VN	VN
XP_002940329	Od	VN	VN	VN	VN
XP_002942058	Od	VN	Ph	Ph	Ph
XP_002941773	Od	mG	Ph	Ph	Ph
XP_002943912	Od	VN or Ph	Ph	Ph	Ph

Table A.2.: Frequently misclassified sequences common to all four data transformations for subtypes Ph and Od. It includes the following columns: GPCRDB identifier, GPCRDB true class, and predicted class for, in turn, the AAC, Digram, ACC and PDBT transformations