Master in Artificial Intelligence (UPC-URV-UB)

Master of Science Thesis

Class C GPCR Metabotropic Glutamate Receptor subtype discrimination using Computational Intelligence methods

Christiana Halka

Advisors: Àngela Nebot, *PhD*, Alfredo Vellido, *PhD*, Francisco Mugica, *PhD*

November 16, 2014

Acknowledgements

This master thesis has been supported by and is part of a broader research effort substantiated as project TIN2012-31377: “KAPPA AIM: Knowledge Acquisition in Pharmacoproteomics using Advanced Artificial Intelligence Methods”, led by Dr. Alfredo Vellido and funded by the Spanish Ministry of Economy and Competitivity.

This thesis would not have been successful without the guidance and support of several individuals who in one way or another contributed to the preparation and completion of this study.

First and foremost, I would like to express my sincere gratitude to my advisors, Dr. Àngela Nebot, Dr. Francisco Mugica and Dr. Alfredo Vellido, for their unfailing support and encouragement throughout my master studies and during the elaboration of my master thesis.

I would also like to extend a very special thanks to fellow researcher Martha Ivón Cárdenas, for her contribution to this master thesis.

Last but not least, I would like to express my love to my family and friends for their support and understanding throughout the duration of my studies.

Abstract

G Protein-Coupled Receptors (GPCRs) are cell membrane proteins with a very relevant role in many biological processes. They have been extensively investigated over the last decade, mostly due to their expected impact in the field of pharmacology. Class C GPCRs, in particular, regulate a number of important physiological functions and are thus intensively pursued as drug targets. In this thesis, GPCRs in general and Metabotropic Glutamate Receptors (mGluR), a key subtype of class C GPCRs, in particular were analyzed using Computational Intelligence techniques. The 3-D structure of most GPCRs is unknown, and, as a result, their functionality is most often investigated through their primary amino acid sequences. An unsupervised Computational Intelligence clustering method, namely Fuzzy c-Means, was used for protein homology detection and discrimination of the different subtypes of Class C GPCRs and mGluR from different transformations of the amino acid sequences based on their physicochemical properties. Fuzzy c-Means was compared to the standard K-Means algorithm in two different settings: The first assumed that the fixed number of clusters is the same as the known receptor subtypes, while the second removed this constraint and focused on an analysis of the stability of the clustering results.

Keywords. G Protein-Coupled Receptors; Metabotropic Glutamate Receptors; Amino Acid Sequence Transformation; Fuzzy c-Means; cluster stability analysis; Cramér's V index.

Contents

1.Introduction 14

2.G Protein-Coupled Receptors 17
2.1 Introduction 17
2.2 GPCRs 18
2.2.1 Structure and Functions 18
2.2.2 Signaling 20
2.2.3 GPCR Classification 23
2.3 Class C GPCRs 24
2.3.1 Representative family members 25
2.3.2 Structure of class C GPCRs 26
2.3.3 Metabotropic Glutamate Receptors (mGluR) 28

3.Related Work 31

4.Materials and Methods 34
4.1 Materials 34
4.2 Methods 36
4.2.1 Alignment-Free Data Transformations 36
4.2.2 Data clustering 38
4.2.2.1 Crisp and Fuzzy partitions 39
4.2.3 K-means Clustering 42
4.2.4 Fuzzy c-Means Clustering 43
4.2.4.1 The Fuzzy c-Means Algorithm 44
4.2.4.2 Parameters of the FCM Algorithm 45
4.2.5 Assessment of stability in fuzzy clustering 48

5.Experimental Study 52
5.1 Experiments with a Fixed Number of Clusters 54
5.1.1 Fuzzy c-Means for class C GPCRs 55
5.1.2 K-means for class C GPCRs 63
5.2 Experiments with varying number of clusters: Cluster Stability Analysis 67
5.2.1 Full Separation Concordance (SeCo) map 69
5.2.1.1 Class C GPCR results 69

5.2.1.2 mGluR results....73
5.2.2 Thresholding the objective function 76
5.2.2.1 Class C GPCR results....77
5.2.2.2 mGluR results....80

6.Conclusions and Future Work....84
6.1 Conclusions....84
6.2 Future Work....86

Bibliography....87

Appendix A....92
Appendix B....96

List of Figures

2.1 Structural diversity among GPCRs: From left to right, class A, class B and class C [60]. ……19

2.2 Two-dimensional generic GPCR.structure. IL1 – IL3: intracellular loops 1-3, EL-1 – EL3: extracellular loops 1-3, Gα: Aplha subunit of a heterotrimeric G-protein, Gβγ: G-beta/gamma heterodimer of heterotrimeric G protein, PKC: Protein kinase C, PKA: Protein Kinase A, GRK: GPCR kinase [59]…….....……20

2.3 Signal transduction by activation/inactivation of heterotrimeric G proteins through GPCR. The subunits of heterotrimeric G proteins (Gα and Gβγ) in their inactivated state are associated with each other. (a) Inactivate State: In inactivation state the GDP is bound to Gα (Gα-GDP). (b) Activate State: In signal transduction, first the GPCR gets activated by changing its conformation which resulted from binding of agonist/ligands to the extracellular region of GPCR. This activated GPCR further activate the inactive G protein to active G protein complex by dissociating the Gα from Gβγ. In active state the GTP is bound to Gα (Gα-GTP). Now free Gα and Gβγ have their own effectors, E1 and E2, respectively, to further transmit the signals and initiate unique intracellular signaling responses. (c) After the signal transduction, the Gα-GTPase activity hydrolyze the bound GTP (Gα-GTP) to GDP and Pi and inactivate the G protein complex by reassociating the Gα with Gβγ. In this state again GDP is bound to Gα (Gα-GDP) in the G protein complex. In this way the activation and inactivation cycle is completed [15]……..22

2.4 Structure of class C GPCRs. (A) Structural organization of class C GPCRs: They have a common structure consisting of a VFT with two lobes (lobe 1 and lobe 2) separating by a cleft as orthosteric site, a 7TM and a CRD for all but $GABA_B$ receptor. (B) Representation of two prototypical class C GPCRs as heterodimer ($GABA_B$ receptor), or homodimer (mGluR). For GABAB receptor, the VFT is directly linked to the 7TM. Two subunits, $GABA_{B1}$ and $GABA_{B2}$, form an obligatory heterodimer. $GABA_{B1}$ is responsible for endogenous ligands binding, while GABAB2 is responsible for G protein activating. For mGluR, the VFT connects to the 7TM via CRD. mGluR form homodimers which could offer two orthosteric sites per dimer. (C) The first solved structure is the VFT of mGlu1 receptor, which shows that the VFT oscillates between an open and a closed conformation [18]…..27

2.5 The three mGluR subtypes [47]……29

2.6 Summary of the roles of mGlu receptors in peripheral tissues [47]……30

4.1 Illustration of clusters of different shapes and dimensions……38

4.2 Representation of data set Z in a two dimensional space……41

4.3 Different distance norms used in fuzzy clustering……...48

5.1 Class specificity for each cluster of the complete C GPCR data set with the AAC transformation……..58

5.2 Class specificity for each cluster of the complete C GPCR data set with the ACC transformation……..59

5.3 Class specificity for each cluster of the complete C GPCR data set with the Digram transformation……..59

5.4 Class specificity for each cluster of the mGluR data set with the AAC transformation ……61

5.5 Class specificity for each cluster of the mGluR data set with the ACC transformation ……61

5 6 Class specificity for each cluster of the mGluR data set with the Digram transformation ……62

5.7 Class specificity for each cluster of the complete C GPCR data set with the AAC transformation……..63

5.8 Class specificity for each cluster of the complete C GPCR data set with the ACC transformation……..63

5.9 Class specificity for each cluster of the complete C GPCR data set with the Digram transformation…….64

5.10 Class specificity for each cluster of the mGluR data set with the AAC transformation ……65

5.11 Class specificity for each cluster of the mGluR data set with the ACC transformation ……66

5.12 Class specificity for each cluster of the mGluR data set with the Digram transformation……..66

5.13 Separation Corcondance map for CGPCR_AAC dataset. ΔSSQ on the y-axis and the median Cramérs' V on the x-axis, for 500 initializations for each value of c, from 2 to 12, for K-Means……..69

5.14 Separation Corcondance map for CGPCR_AAC dataset. Representation as in previous figure; (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index……..69

5.15 Separation Corcondance map for CGPCR_ACC dataset. ΔSSQ on the y-axis and the median Cramérs' V on the x-axis, for 500 initializations for each value of c, from 2 to 12, for K-Means……..70

5.16 Separation Corcondance map for CGPCR_ACC dataset. Representation as in previous figure; (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index……..70

5.17 Separation Corcondance map for CGPCR_Digram dataset. ΔSSQ on the y-axis and the median Cramérs' V on the x-axis, for 500 initializations for each value of c, from 2 to 12, for K-Means……...71

5.18 Separation Corcondance map for CGPCR_Digram dataset. Representation as in previous figure; (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index……..71

5.19 Separation Corcondance map for mGluR_AAC dataset. ΔSSQ on the y-axis and the median Cramérs' V on the x-axis, for 500 initializations for each value of c, from 2 to 10, for K-Mean..……73

5.20 Separation Corcondance map for mGluR_AAC dataset. Representation as in previous figure; (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index……..……73

5.21 Separation Corcondance map for mGluR_ACC dataset. ΔSSQ on the y-axis and the median Cramérs' V on the x-axis, for 500 initializations for each value of c, from 2 to 10, for K-Means……..…..……74

5.22 Separation Corcondance map for mGluR_ACC dataset. Representation as in previous figure; (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index……..……74

5.23 Separation Corcondance map for mGluR_ Digram dataset. ΔSSQ on the y-axis and the median Cramérs' V on the x-axis, for 500 initializations for each value of c, from 2 to 10, for K-Means……..……...75

5.24 Separation Corcondance map for mGluR_Digram dataset. Representation as in previous figure; (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index……..……..75

5.25 Separation Corcondance map for CGPCR_AAC dataset using 10% threshold for separation metric. K-Means. Representation as in previous figures ……..……77

5.26 Separation Corcondance map for CGPCR_AAC dataset using 10% threshold for separation metric. (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index. Representation as in previous figures…..…….……77

5.27 Separation Corcondance map for CGPCR_ACC dataset using 10% threshold for separation metric. K-Means. Representation as in previous figures……..……78

5.28 Separation Corcondance map for CGPCR_ACC dataset using 10% threshold for separation metric. (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index. Representation as in previous figures……..…….78

5.29 Separation Corcondance map for CGPCR_ Digram dataset using 10% threshold for separation metric. K-Means. Representation as in previous figures……..……..…79

5.30 Separation Corcondance map for CGPCR_Digram dataset using 10% threshold for separation metric. (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index. Representation as in previous figures.……..……79

5.31 Separation Corcondance map for mGluR_AAC dataset using 10% threshold for separation metric. K-Means. Representation as in previous figures……..……..80

5.32 Separation Corcondance map for mGluR_AAC dataset using 10% threshold for separation metric. (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index. Representation as in previous figures.……..……80

5.33 Separation Corcondance map for mGluR_ACC dataset using 10% threshold for separation metric. K-Means. Representation as in previous figures……..……..81

5.34 Separation Corcondance map for mGluR_ACC dataset using 10% threshold for separation metric. (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index. Representation as in previous figures.……..……..81

5.35 Separation Corcondance map for mGluR_ Digram dataset using 10% threshold for separation metric. K-Means. Representation as in previous figures……..……82

5.36 Separation Corcondance map for mGluR_ Digram dataset using 10% threshold for separation metric. (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index. Representation as in previous figures.……..……82

B.1 CGPCR_AAC dataset's histogram.……97

B.2 CGPCR_ACC dataset's histogram.……97

B.3 CGPCR_Digram dataset's histogram.……...98

B.4 mGluR_AAC dataset's histogram.……..98

B.5 mGluR_ACC dataset's histogram.……99

B.6 mGluR_Digram dataset's histogram.……99

B.7 The Accuracy and the number of rejected cases compared with the threshold values for CGPCR_AAC dataset……..……102

B.8 The Accuracy and the number of rejected cases compared with the threshold values for CGPCR_ACC dataset……102

B.9 The Accuracy and the number of rejected cases compared with the threshold values for CGPCR_Digram dataset……103

B.10 The Accuracy and the number of rejected cases compared with the threshold values for mGluR_AAC dataset……..103

B.11 The Accuracy and the number of rejected cases compared with the threshold values for mGluR_ACC dataset……..104

B.12 The Accuracy and the number of rejected cases compared with the threshold values for mGluR_Digram dataset……..104

List of Tables

2.1 G Protein-Coupled Receptor Families……..23

4.1The five major classes of the GPCR superfamily according to GPCRDB ……35

4.2 The main eight subtypes of metabotropic glutamate receptors grouped into three categories ……36

5.1 List of the six datasets used in this experiments. The three sequence transformations for the class C GPCR, plus the three sequence transformation for mGluR ……...53

5.2 Fuzziness Parameter *m* for each data set. In this table, the upper bound of the fuzziness parameter *m* for each data set is displayed, as well as the chosen value of m for each one of them. *N* is the number of cases in the dataset; *p* the number of attributes; and *c* the number of clusters……..57

5.3 Entropy measure for the complete C GPCR data set with the AAC transformation….58

5.4 Entropy measure for the complete C GPCR data set with the ACC transformation….59

5.5 Entropy measure for the complete C GPCR data set with the Digram transformation.60

5.6 Entropy measure for the mGluR data set with the AAC transformation……..61

5.7 Entropy measure for the mGluR data set with the ACC transformation……..61

5.8 Entropy measure for the mGluR data set with the Digram transformation…….62

5.9 Entropy measure for the complete C GPCR data set with the AAC transformation…..63

5.10 Entropy measure for the complete C GPCR data set with the ACC transformation...64

5.11Entropy measure for the complete C GPCR data set with the Digram transformation64

5.12 Entropy measure for the mGluR data set with the AAC transformation……66

5.13 Entropy measure for the mGluR data set with the ACC transformation……66

5.14 Entropy measure for the mGluR data set with the Digram transformation……...66

A.1 Class specificity in each cluster of CGPCR_AAC dataset..……92

A.2 Class specificity in each cluster of CGPCR_ACC dataset...……92

A.3 Class specificity in each cluster of CGPCR_Digram dataset..……93

A.4 Class specificity in each cluster of mGluR_AAC dataset.....…….93

A.5 Class specificity in each cluster of mGluR_ACC dataset.....…….93

A.6 Class specificity in each cluster of mGluR_Digram dataset...……..93

A.7 Class specificity in each cluster of CGPCR_AAC dataset...…….94

A.8 Class specificity in each cluster of CGPCR_ACC dataset...…….94

A.9 Class specificity in each cluster of CGPCR_Digram dataset...……95

A.10 Class specificity in each cluster of mGluR_AAC dataset...…….95

A.11 Class specificity in each cluster of mGluR_ACC dataset...…….95

A.12 Class specificity in each cluster of mGluR_Digram dataset...……..……..95

B.13 Membership Value for the CGPCR_AAC dataset……..……..105

B.14 Membership Value for the CGPCR_ACC dataset……..……..105

B.15 Membership Value for the CGPCR_Digram dataset……..……..106

B.16 Membership Value for the mGluR_AAC dataset...……..…….106

B.17 Membership Value for the mGluR_ACC dataset…..……..……..107

B.18 Membership Value for the mGluR_Digram dataset…...……...……..107

B.19 The Accuracy and the number of rejected cases compared with the threshold values for CGPCR_AAC data set…….……108

B.19 The Accuracy and the number of rejected cases compared with the threshold values for CGPCR_ACC data set…….……108

B.19 The Accuracy and the number of rejected cases compared with the threshold values for CGPCR_Digram data set...……..……109

B.19 The Accuracy and the number of rejected cases compared with the threshold values for mGluR_AAC data set……...……109

B.19 The Accuracy and the number of rejected cases compared with the threshold values for mGluR_ACC data set…..……..……...……110

B.19 The Accuracy and the number of rejected cases compared with the threshold values for mGluR_Digram data set……..……..……110

Chapter 1

Introduction

G-Protein Coupled Receptors (GPCRs) are cell membrane proteins with a very important role in regulating many of the cell functions. This is mostly the result of their ability to transmit extracellular signals. Such characteristic makes them a much sought after target in pharmacology. This has led, over the last decade, to very active research in this particular area of the broader subject of proteomics, which can be seen as one more thread of the ongoing “-omics revolution” in biology.

The functionality of a protein depends widely on its 3-D configuration, which determines its ability for certain ligand binding. This compound binding behavior determines the role of the protein in the metabolic pathways: the chains of chemical reactions occurring within a cell. Currently, the 3-D structure of only about 21 GPCRs is fully determined, with the majority of them validated only in the last few years. Only one of them, found in 2014, belongs to class C, which is the object of this thesis. When the information about the 3-D structure is not available, the investigation of the functionality of a protein must often be limited to the analysis of its primary amino acid sequence, in the understanding that the sequential ordering of amino acids at least partially determines the functionality of the receptor.

Unaligned symbolic sequences are not easy to analyze directly, but, recently, alternative approaches using machine learning and computational intelligence techniques for the analysis of alignment-free sequences have been proposed. The current thesis focuses on the alignment-free analysis of class C GPCRs. Here, alignment-free full sequences were used to limit information loss (in contrast with the analysis of aligned sequences, in which the loss of part of the available information is unavoidable). More specifically, three existing alignment-free transformations were used, the amino acid composition transformation, the auto cross covariance transformation and the digram transformation, which will be described in detail in the following chapters.

The class C of GPCRs has become an increasingly important target for new

therapies, particularly in areas such as Fragile-X syndrome, schizophrenia, Alzheimer's disease, Parkinson's disease, epilepsy, L-DOPA-induced dyskinesias, generalized anxiety disorder, migraine, chronic pain, gastroesophageal reflux disorder, hyperparathyroidism and osteoporosis. For this reason, the interest of their study in pharmacology is self-explanatory. Moreover, metabotropic glutamate receptors (mGluR) in particular, a subtype of the class C of GPCRs are known to play an important neuro-modulatory role throughout the brain. They are in fact targets for therapeutic intervention for a number of psychiatric and neurological disorders. Thus the investigation of class C and its subtypes, particularly mGluR, is deemed to be truly important.

A Computational Intelligence technique, namely the Fuzzy c-Means (FCM) algorithm, was used for the analysis and discrimination of class C GPCRs and mGluRs. FCM is an unsupervised fuzzy clustering technique in which data points are not bound to belong to a single cluster and can in fact belong to more than one with different degrees of membership. The fuzzy membership values of each data point that FCM provides provide us with information about the strength of the association between a data point and a particular data cluster. Thus, information about the proteins homology and the level of association of each protein sequence with each cluster can be extracted. Moreover, FCM can infer the characteristics of each subtype of class C GPCRs or mGluR; for example, whether a single compact group (cluster) for each subtype exists, or if some subtypes overlap with other, or with several others.

FCM is an extension of K-Means, a stalwart method for data clustering, successfully in use for decades. K-Means is based on crisp cluster assignments and its limitations are well-studied and include the lack of a closed criterion for the choice of the number of clusters *K* and the fact that, under different initializations, the algorithm may yield very different solutions. Recent experimental evidence has shown that K-Means solutions that might be expected to be similar according to the final value of the objective function may in fact be quite dissimilar. This suggests the convenience of using the objective function as a criterion of model optimality *only in combination with some cluster stability criterion* in order to achieve cluster partition reproducibility. One such combined criterion is the Separation and Concordance (SeCo) map, which, in this thesis, we extend to FCM by first defining weighted contingency tables and a corresponding weighted Cramér's V index.

The current thesis is organized as follows: Chapter 2 focuses on a general description of GPCRs and their biological role. A non-exhaustive overview of the basic characteristics, such as structure, functions, ligand bindings and classification, of GPCRs, Class C of GPCRs and mGluR are presented.

In Chapter 3, some work of interest published in this particular field and related to our research is briefly reviewed.

Chapter 4 contains general information about the materials and the methods used for the purposes of this research. Fuzzy clustering and, specifically, the FCM algorithm are summarily described. This chapter also includes a brief introduction to the K-Means algorithm, the Separation/Concordance (SeCo) maps and their proposed novel extension to fuzzy and probabilistic clustering methods.

Chapter 5 contains a summary report of the experiments carried out, their results and discussion.

Finally, Chapter 6 wraps up the thesis with some conclusions and an outline of possible future lines of work.

Further experiments related to the research reported in chapter 5 are compiled in appendices A and B as supplementary material.

Chapter 2

G Protein-Coupled Receptors

2.1 Introduction 17

2.2 GPCRs 18

2.2.1 Structure and Functions 18

2.2.2 Signaling 20

2.2.3 Classification 23

2.3 Class C GPCRs 24

2.3.1 Representative Family Members 25

2.3.2 Structure of class C GPCRs 26

2.3.3 Metabotropic Glutamate Receptors 28

2.1 Introduction

G-protein-coupled receptors (GPCRs) constitute a large and diverse family of proteins whose primary functions include the transduction of extracellular stimuli into intracellular signals (that is, they work as “signal gatekeepers” in the cell membrane). They sense molecules outside the cell and, as a result, activate internal signal transduction pathways, and consequently, cellular responses.

These receptors are among the largest and most diverse protein families in mammals. They are known to consist of seven membrane-spanning helices, an extracellular N-terminus and an intracellular C-terminus. For this reason, they are also known as 7-TransMembrane (7-TM) receptors or heptahelical receptors [1].

GPCRs can be found in eukaryote cells, including yeast, choanoflagellates (unicellular precursors of animals) and animals. The diversity of GPCRs is dictated both by the multiplicity of stimuli to which they respond and by the variety of intracellular signaling pathways they activate [1]. The ligands that bind and activate these receptors include light-sensitive compounds, odors, pheromones, hormones and neurotransmitters, and vary in size from small molecules to peptides and to large proteins.

G proteins were discovered when Alfred G. Gilman and Martin Rodbell investigated stimulation of cells by adrenaline. They found that, when adrenaline binds to a receptor, the receptor does not stimulate enzymes directly. Instead, the receptor stimulates a G protein, which stimulates an enzyme. An example is adenylate cyclase, which produces the second messenger cyclic AMP [5]. For this discovery, they were awarded the 1994 Nobel Prize in Physiology or Medicine [6].

2.2 GPCRs

2.2.1 Structure and Functions

All GPCRs are characterized by an extracellular N-terminus, followed by seven TM (7-TM) α-helices (TM-1 to TM-7) connected by three intracellular (IL-1 to IL-3) and three extracellular loops (EL-1 to EL-3), and finally an intracellular carboxyl terminus (C-terminus), as shown in Figure 2.2.

They share two features: the presence of 7 TM α-helices in the receptor protein, as previously mentioned and also the ability to couple to heterotrimeric G-proteins, which can increase or reduce the activity of effector enzymes, like phospholipase C and adenylate cyclase [2]. The most variable structures among the family of GPCRs are the carboxyl terminus, the intracellular loop spanning TM5 and TM6, and the amino terminus. The greatest diversity is observed in the amino terminus. This sequence is relatively short (10–50 amino acids) for monoamine and peptide receptors, and much larger (350–600 amino acids) for glycoprotein hormone receptors, and the glutamate family receptors. The largest amino terminal domains are observed in the adhesion family receptors [16]. The structure diversity among GPCRs is illustrated in Figure 2.1.

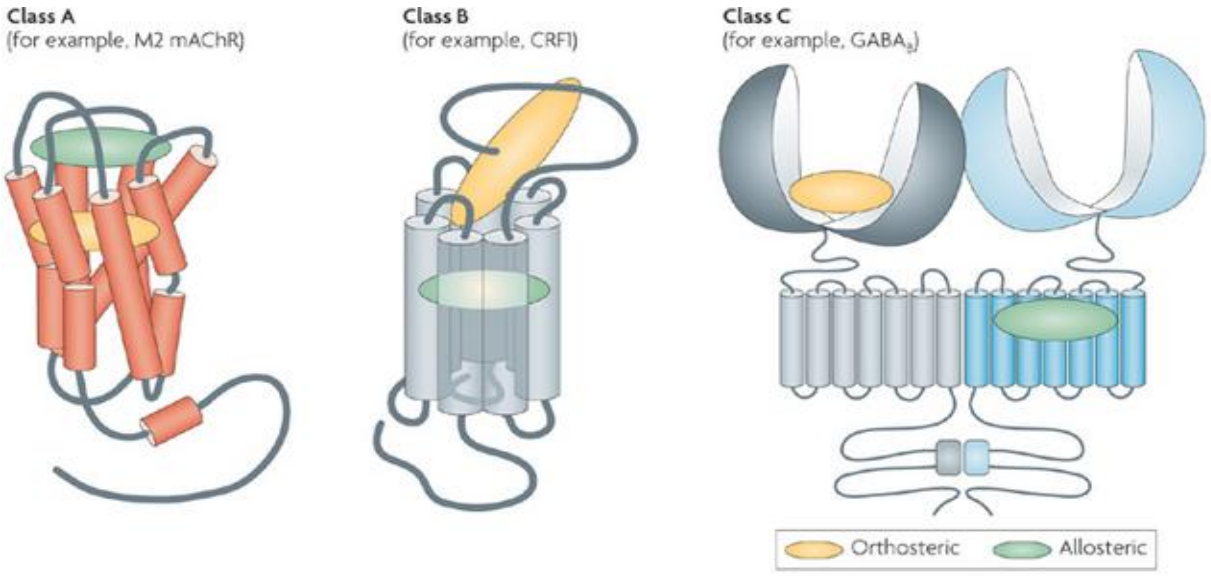

Figure 2.1: Structural diversity among GPCRs: From left to right, class A, class B and class C [60].

The GPCR arranges itself into a tertiary (3-D) structure resembling a barrel, with the seven TM helices forming a cavity within the plasma membrane that serves a ligand-binding domain that is often covered by EL-2. Ligands may also bind elsewhere, however, as is the case for bulkier ligands, for example proteins or large peptides, which instead interact with the extracellular loops, or, as illustrated by the class C mGluR subtype, the N-terminal tail.

The class C of GPCR, in particular, is distinguished by their large N-terminal tail, which also contains a ligand-binding domain. Upon glutamate-binding to an mGluR, the N-terminal tail undergoes a conformational change that leads to its interaction with the residues of the extracellular loops and TM domains. The eventual effect of all three types of agonist-induced activation is a change in the relative orientations of the TM helices leading to a wider intracellular surface and "revelation" of residues of the intracellular helices and TM domains crucial to signal transduction function (i.e., G-protein coupling). Inverse agonists and antagonists may also bind to a number of different sites, but the eventual effect must be prevention of this TM helix reorientation.

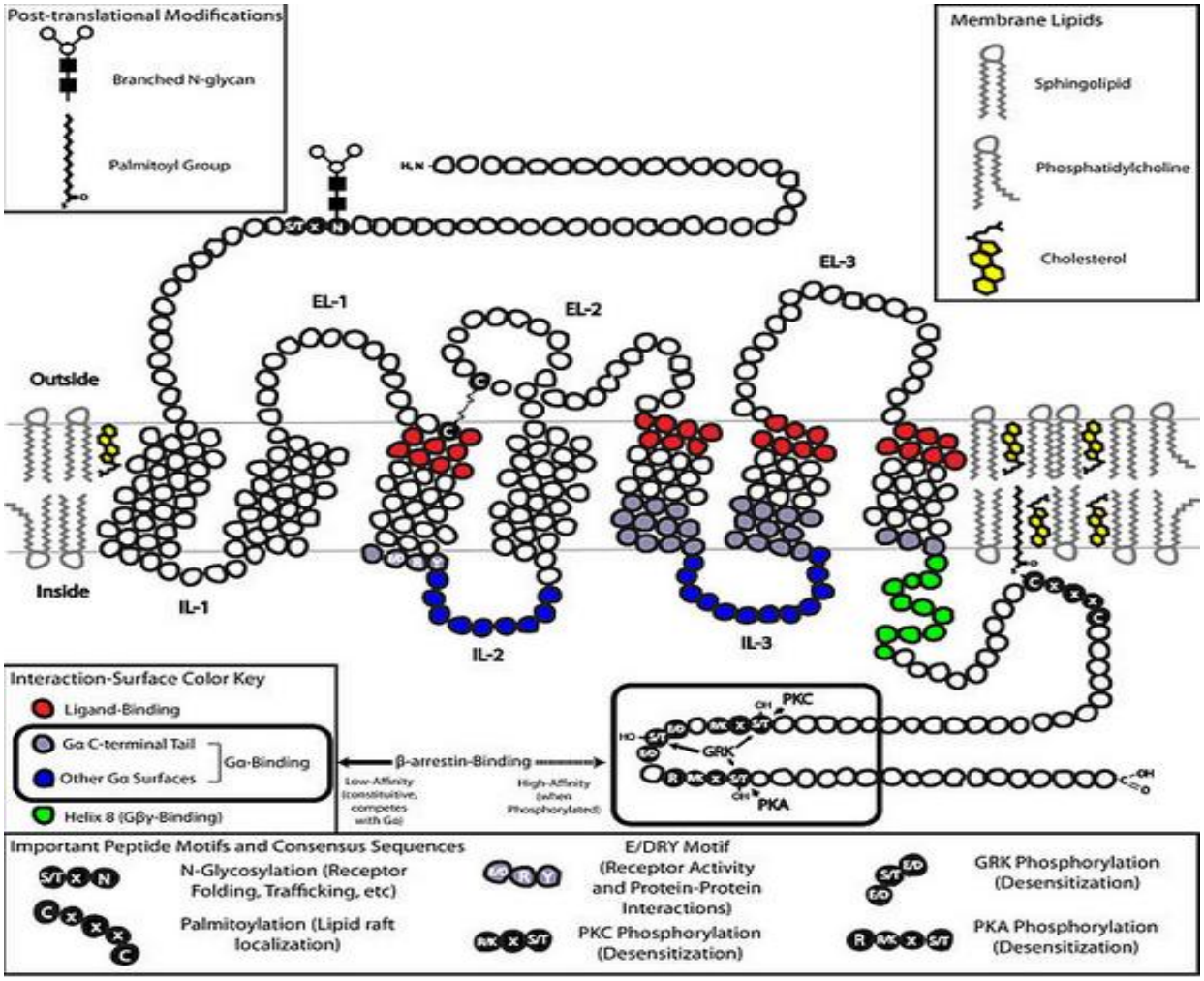

Figure 2.2: Two-dimensional generic GPCR.structure. IL1 – IL3: intracellular loops 1-3, EL-1 – EL3: extracellular loops 1-3, Gα: Aplha subunit of a heterotrimeric G-protein, Gβγ: G-beta/gamma heterodimer of heterotrimeric G protein, PKC: Protein kinase C, PKA: Protein Kinase A, GRK: GPCR kinase [59].

GPCRs respond to extracellular signals mediated by a huge diversity of agonists, ranging from proteins to biogenic amines to protons, but all transduce this signal via a mechanism of G-protein coupling. This is made possible by a guanine-nucleotide exchange factor (GEF) domain primarily formed by a combination of IL-2 and IL-3 along with adjacent residues of the associated TM helices.

2.2.2 Signaling

Cell signaling is one of the important processes required for the normal growth and development of the cell. The basic cell signaling machinery involves a receptor molecule that perceives the signal. The signal or primary stimulus that activates these receptors could be light, hormone, odorant, antigen, neurotransmitter or the surface of another cell,

which convey into the cell via the membrane receptor, through signal transduction triad (receptor/transducer/effector) [13]. The second messenger could be Ca2+ (for ion channels) cAMP and cGMP (for adenylyl and guanlyl cyclases), inositol-1, 4,5-triphosphate (IP3), diacyl glycerol (DAG) and arachidonic acid (for phospholipases). The triad is responsible for converting the signal from first to second messenger, which could be further regulated by protein kinases or phosphatases in the cytoplasm. The target of the signal may be enzymes, intracellular receptors, special transport vehicles and finally transcription factors, which ultimately controls the gene expression [14].

The cell has several signaling mechanisms, but a very important signaling cascade is formed by GTP binding proteins, or G proteins for short. One more molecule that is involved in this signaling cascade and forms an important part of the cascade is the GPCR. It is known that the signals are mostly perceived at the level of membrane and therefore TM events are the likely routes for signal generation and transduction [15].

Heterotrimeric G proteins (Gα, Gβ/Gγ subunits) constitute one of the most important components of cell signaling cascade. GPCRs perceive many extracellular signals and transduce them to heterotrimeric G proteins, which further transduce these signals intracellularly to appropriate downstream effectors and thereby play an important role in various signaling pathways.

The activation/inactivation cycle of G protein through GPCR is shown in Figure 2.3. In the inactive state, Gα is bound to Gβγ dimer and GDP, Figure 2.3 (a). G protein mediated signaling starts by binding of an agonist molecule that leads to activation of GPCR. GPCR is also a guanine nucleotide exchange factor that promotes the exchange of guanosine disphosphate (GDP)/guanosine triphosphate (GTP) associated with the Gα subunit. Therefore, the activated GPCR catalyzes exchange of GTP for GDP on the Gα subunit, as a result conformational changes takes place in the GPCR, which leads to dissociation of Gβγ dimer from Gα and thus activates multiple molecules of G proteins, Figure 2.3 (b). The G proteins activated in this way constitute an amplified representation of the activated GPCR. Activated Gα and Gβγ proteins in turn binds to various effectors and thereby switches it either on or off in different systems, and effectors continue to pass the signal to different kinds of second messengers. Here intrinsic GTPase activity of Gα

comes into play, that leads to conversion of bound GTP into GDP and hence the inactivation of G proteins cascade, Figure 2.3(c). GTPase activity of the Gα subunits may also be regulated by regulators of G proteins signaling (RGS proteins) as well as effectors. Moreover, effector enzymes such as adenylyl cyclases may also regulate the activation of G proteins by receptors [15].

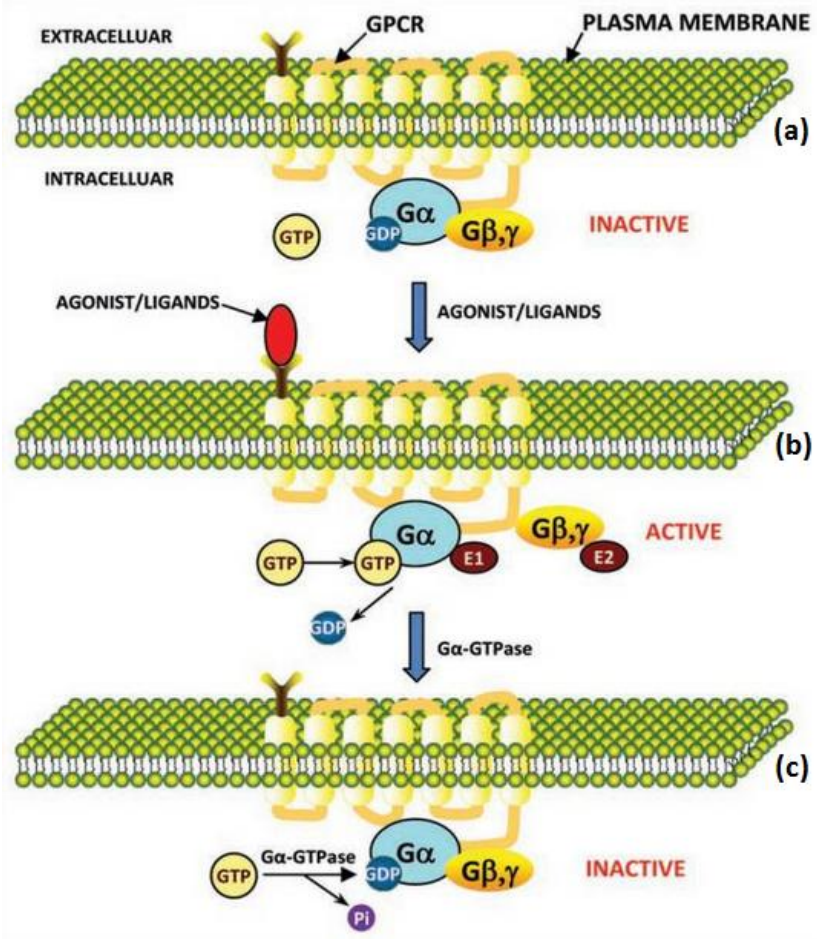

Figure 2.3**:** Signal transduction by activation/inactivation of heterotrimeric G proteins through GPCR. The subunits of heterotrimeric G proteins (Gα and Gβγ) in their inactivated state are associated with each other. **(a)** Inactivate State: In inactivation state the GDP is bound to Gα (Gα-GDP). **(b)** Activate State: In signal transduction, first the GPCR gets activated by changing its conformation which resulted from binding of agonist/ligands to the extracellular region of GPCR. This activated GPCR further activate the inactive G protein to active G protein complex by dissociating the Gα from Gβγ. In active state the GTP is bound to Gα (Gα-GTP). Now free Gα and Gβγ have their own effectors, E1 and E2, respectively, to further transmit the signals and initiate unique intracellular signaling responses. **(c)** After the signal transduction, the Gα-GTPase activity hydrolyze the bound GTP (Gα-GTP) to GDP and Pi and inactivate the G protein complex by reassociating the Gα with Gβγ. In this state again GDP is bound to Gα (Gα-GDP) in the G protein complex. In this way the activation and inactivation cycle is completed [15].

Although they are classically understood to work only together, GPCRs may signal through G-protein-independent mechanisms, and heterotrimeric G-proteins may play functional roles independent of GPCRs. Moreover, from recent studies it appears that heterotrimeric G-proteins may also take part in non-GPCR signaling.

2.2.3 GPCR Classification

As previously mentioned, the members of the GPCRs superfamily are diverse in their primary structure, and this has been used for the phylogenetic classification of the family members. Attwood and Findlay were the first in trying to classify the GPCR family, in 1993. They developed sequenced-based fingerprints of the seven characteristic GPCR hydrophobic domains and these were used as diagnostic tools for identifying sequences belonging to the GPCR family [7].

A more comprehensive view of the human GPCR repertoire was possible when the first draft of the human genome became available, in 2001 [8][9]. Nearly 800 different human genes have been predicted from genome sequence analysis. Although numerous classification schemes have been proposed, the GPCR superfamily was classically divided into three main classes (A, B, and C) with no detectable shared sequence homology between classes.

In 2006, Fredriksson and colleagues proposed an alternative classification system called GRAFS (Glutamate, Rhodopsin, Adhesion, Frizzled/Taste2, Secretin) [10]. According to this system, GPCRs could be grouped into 6 classes based on sequence homology and functional similarity, as shown in Table 2.1.

GPCR classes	Description
Class A (or 1)	Rhodopsin-like receptors
Class B (or 2)	Secretin receptor family
Class C (or 3)	Metabotropic glutamate receptors
Class D (or 4)	Fungal mating pheromone receptors
Class E (or 5)	Cyclic AMP receptors
Class F (or 6)	Frizzled/Smoothened receptors

Table 2.1: G Protein-Coupled Receptor Families

The rhodopsin family is by far the largest and most diverse of these families, forming four main groups with 13 sub-branches, and the members are characterized by preserved sequence motifs and are thought to share similar activation mechanisms [11]. It was shown that rhodopsin is a membrane-spanning protein that has the ability to transfer energy from light into intracellular signaling cascades, an ability that allows us to see [4].

GPCRs are major drug targets, and are consequently the subject of considerable research interest. It has been reported that the repertoire of GPCRs for endogenous ligands consists of approximately 400 receptors in humans and mice [12]. Most GPCRs are identified on the basis of their DNA sequences, rather than the ligand they bind, those that are unmatched to known natural ligands are designated by as orphan GPCRs, or unclassified GPCRs [17].

2.3 Class C GPCRs

Class C GPCRs represent a distinct group of the GPCR superfamily. Among the other families, class C GPCRs are defined by two unique structural features: first, they possess a large extracellular domain that is distal to the 7TM and second, they form constitutive dimmers with unique activation modes compared with other classes of GPCRs [18]. Structurally, they consist of four elements: an N-terminal signal sequence, a large hydrophilic extracellular agonist-binding region containing several conserved cystein residues which could be involved in disulphide bonds, a shorter region containing 7-TM domains, and a C-terminal cytoplasmic domain of variable length.

Class C GPCRs can be further subdivided into metabotropic glutamate receptors (mGlu receptors), *γ*-aminobutyric $acid_B$ receptors ($GABA_B$ receptors), calcium-sensing receptors (CaSR receptors), sweet and amino acid taste receptors, pheromone receptors, odorant receptors in fish and several orphan receptors [19]. Class C GPCRs are also involved in important physiological processes throughout the body: mGluR, GABAB, and CaSR represent an important new type of therapeutic targets that are integral to disorders that affect the central neural system (CNS) and calcium homeostasis [20]. The taste receptors, on the other hand, attract significant attention from food companies because the taste additives that target these receptors represent a key feature of the large food industry market [21].

2.3.1 Representative family members

Metabotropic glutamate receptors are localized almost exclusively in the mammalian central neural system (CNS), and they participate in the modulation of synaptic transmission (in majority excitatory synapses) and neuronal excitability. The mGluR are, in turn, sub-divided into eight subtypes ($mGlu_1$ - $mGlu_8$). The eight mGluR subtypes estimated to date can be grouped into three subgroups based on amino acid sequence similarity, agonist pharmacology and G-protein coupling property: *Group I* being comprised of $mGlu_1$ and $mGlu_5$; *Group II* of $mGlu_2$ and $mGlu_3$; and *Group III* containing the remaining four subtypes namely $mGlu_4$, $mGlu_6$, $mGlu_7$ and $mGlu_8$ [20].

The mGlu receptors are important contributors to the synaptic transmission of the major excitatory neurotransmitter in the body, thus they are attractive drug targets [22]. Recent studies continue to validate the therapeutic utility of mGluR ligands in neurological and psychiatric disorders, such as Parkinson's disease [24], Fragile X syndrome [25], Alzheimer's disease [26], anxiety, and schizophrenia [27].

The GABA receptor is a major inhibitory neurotransmitter in the mammalian CNS. As the metabotropic receptor for GABA, $GABA_B$ receptor mediates slow and prolonged synaptic inhibition. The $GABA_B$ receptors contain two amino-terminal sushi-repeats, which are protein-protein interaction motifs that are expected to serve as an extracellular targeting signal that dictates sub cellular localization. The metabolic signaling of (S)-glutamic acid (Glu) and *γ*-aminobutyric acid (GABA) mediated by these receptors supplement the fast synaptic transmission mediated by families of ligand-gated ion channels for both of these neurotransmitters. Only two $GABA_B$ receptor genes have been identified, $GABA_{B1}$ and $GABA_{B2}$ [22]. In addition to a role in neuronal excitability and plasticity, $GABA_B$ receptor may promote neuron survival under conditions of metabolic stress, ischemia, or apoptosis. This receptor is a promising target for the treatment of many diseases, including spasticity, neuropathic pain, drug addiction, schizophrenia, anxiety, depression and epilepsy.

The CaSR is a unique class C GPCR that can be activated by ions without the cooperation of other ligands. This receptor is highly sensitive to a very slight change in

extracellular Ca^{2+} concentrations, which ensures its significant role in regulating calcium homeostasis. The CaSR is involved in several disorders, including hyperparathyroidism, osteoporosis and different forms of hypocalcemia. The clinical success of the Cinacalcet drug indicates that more efforts should be devoted to the discovery of novel ligands that modulate CaSR receptor activation [18].

Class C GPCRs contain three taste receptor subunits (T1R1, T1R2, and T1R3) that form two heterodimers, the sweet receptor (T1R2/T1R3) and the umami receptor (T1R1/T1R3) [28]. In addition to natural sugars, the sweet taste receptor is also sensitive to artificial sweeteners, sweet proteins and some D-amino acids. In most mammals, the umami receptor can be activated by L-amino acids, whereas the human orthologue is only sensitive to monosodium glutamate and L-aspartate. Flavor enhancers, such as purine nucleotides, have the ability to potentiate umami receptor function. These artificial sweeteners and flavor enhancers represent a large food sector market.

2.3.2 Structure of class C GPCRs

Class C GPCRs are composed of an exceptionally large extracellular domain, a heptahelical TM domain separated by alternating intracellular and extracellular loops (IL1-IL3 and EL1-EL3, respectively) and an intracellular carboxyl-terminal (C-terminal) domain, as shown in Figure 2.4(A). One distinct structural feature of class C GPCRs is the extracellular domain that contains a Venus flytrap (VFT) module and a cysteine rich domain (CRD) whereas it is not present in the $GABA_B$ receptor. The C-terminal tail of class C GPCRs is a highly variable domain and plays a role in scaffolding and signaling protein coupling. All the domains except for the intracellular C-terminal domain provide plentiful ligand action sites. The other unique characteristic of class C GPCRs is their mandatory dimerization, either as homodimers (mGlu and CaS receptors) or heterodimers ($GABA_B$ receptor and T1Rs), as shown in Figure 2.4(B). Among class C GPCRs, the VFT of the mGlu1 receptor is the first for which a crystal structure was solved, Figure 2.4(C) [29].

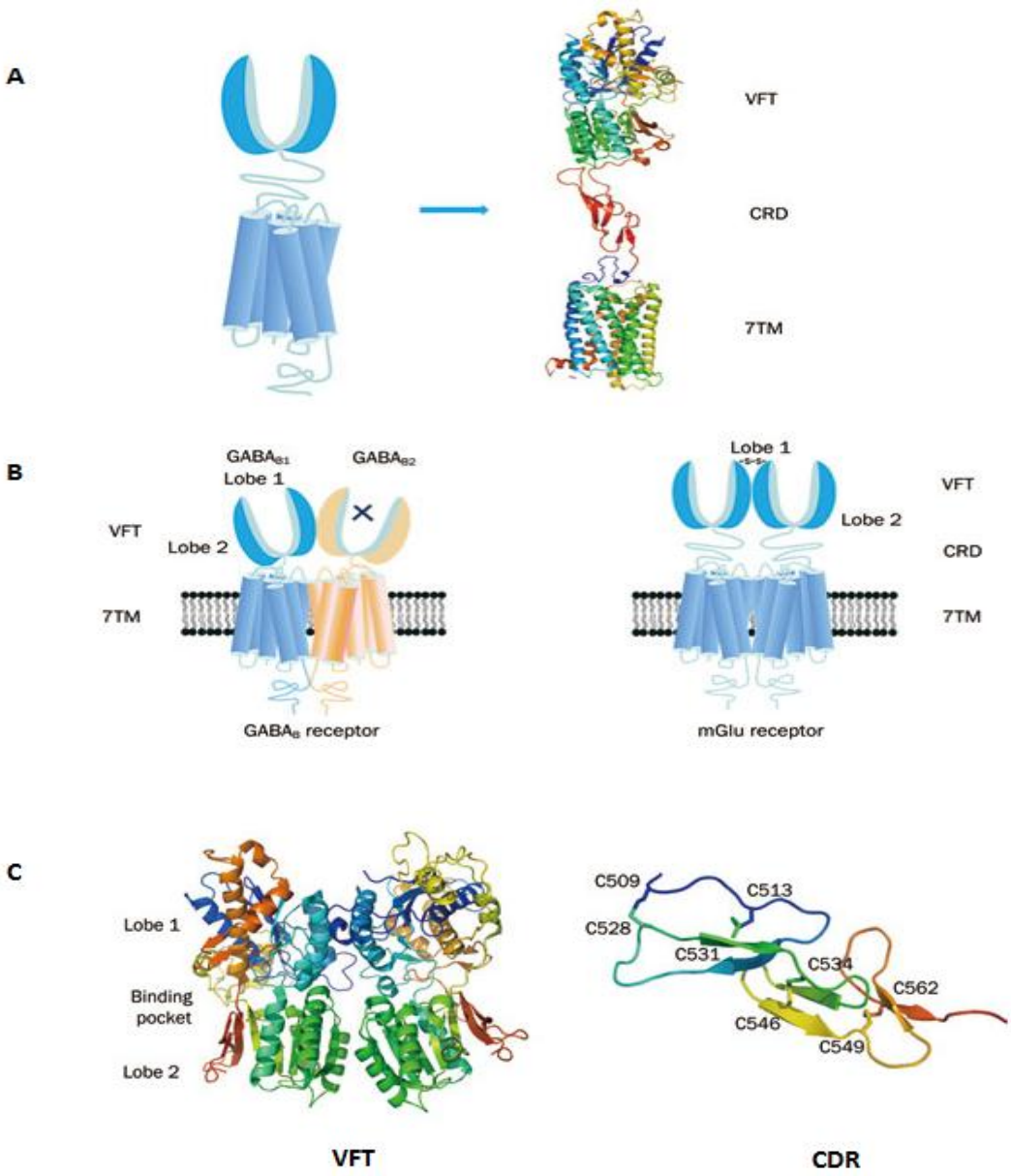

Figure 2.4: Structure of class C GPCRs. (A) Structural organization of class C GPCRs: They have a common structure consisting of a VFT with two lobes (lobe 1 and lobe 2) separating by a cleft as orthosteric site, a 7TM and a CRD for all but $GABA_B$ receptor. (B) Representation of two prototypical class C GPCRs as heterodimer ($GABA_B$ receptor), or homodimer (mGluR). For $GABA_B$ receptor, the VFT is directly linked to the 7TM. Two subunits, $GABA_{B1}$ and $GABA_{B2}$, form an obligatory heterodimer. $GABA_{B1}$ is responsible for endogenous ligands binding, while $GABA_{B2}$ is responsible for G protein activating. For mGluR, the VFT connects to the 7TM via CRD. mGluR form homodimers which could offer two orthosteric sites per dimer. (C) The first solved structure is the VFT of $mGlu_1$ receptor, which shows that the VFT oscillates between an open and a closed conformation [18].

2.3.3 Metabotropic Glutamate Receptors

The metabotropic glutamate receptors, or mGluRs, of special relevance in this thesis, are a type of glutamate receptor that are active through an indirect metabotropic process. They are members of the group C family of G-protein-coupled receptors. Like all glutamate receptors, mGluRs bind with glutamate, an amino acid that functions as an excitatory neurotransmitter.

The mGluRs perform a variety of functions in the central and peripheral nervous systems: For example, they are involved in learning, memory, anxiety, and the perception of pain [50]. They are found in pre- and postsynaptic neurons in synapses of the hippocampus, cerebellum [51], and the cerebral cortex, as well as other parts of the brain and in peripheral tissues [52].

Like other metabotropic receptors, mGluRs have seven transmembrane domains that span the cell membrane [53]. Unlike ionotropic receptors, metabotropic glutamate receptors are not ion channels. Instead, they activate biochemical cascades, leading to the modification of other proteins, as for example ion channels. This can lead to changes in the synapse's excitability, for example by presynaptic inhibition of neurotransmission [8], or modulation and even induction of postsynaptic responses.

There are eight different subtypes of mGluRs, named as $mGluR_1$ to $mGluR_8$. As previously mentioned, the mGluR family is divided into three groups based on amino acid homology, signal transduction pathways and pharmacologic interest, as illustrated in Figure 2.5.

- Group I: $mGluR_1$, $mGluR_5$
- Group II: $mGluR_2$, $mGluR_3$
- Group III: $mGluR_4$, $mGluR_6$, $mGluR_7$, $mGluR_8$

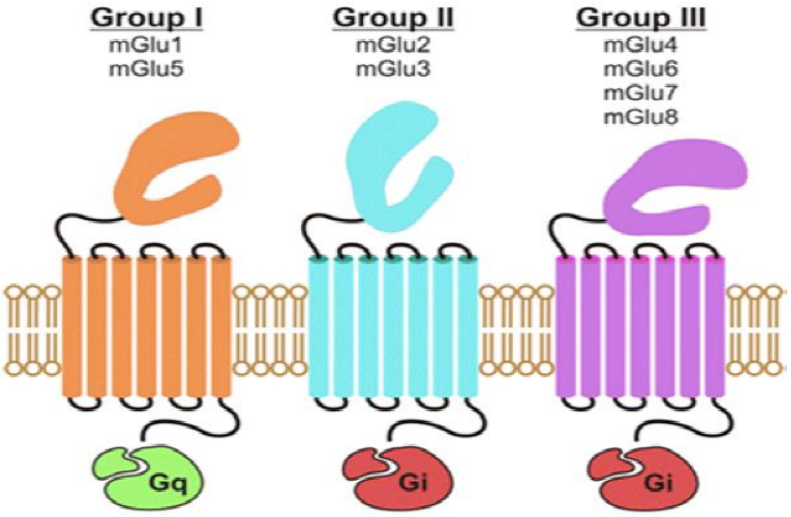

Figure 2.5: The three mGluR subtypes [47].

L-Glutamate is produced by a great variety of the peripheral tissues in both health and disease. Like other components of the glutamatergic system, mGluRs also have a widespread distribution outside the CNS, including cells that do not have a neuronal phenotype (See Figure 2.6).

Analysis of the recent literature reveals an extraordinary potential, particularly for *Group I* and *III* mGluRs in the treatment of peripheral disorders of the most diverse nature, such as endocrine dysregulation, aberrant cell proliferation, and gastrointestinal disorders. The significance of these findings is that pharmacological tools originally designed for mGluRs in the CNS may also be directed toward new disease targets in the periphery [47].

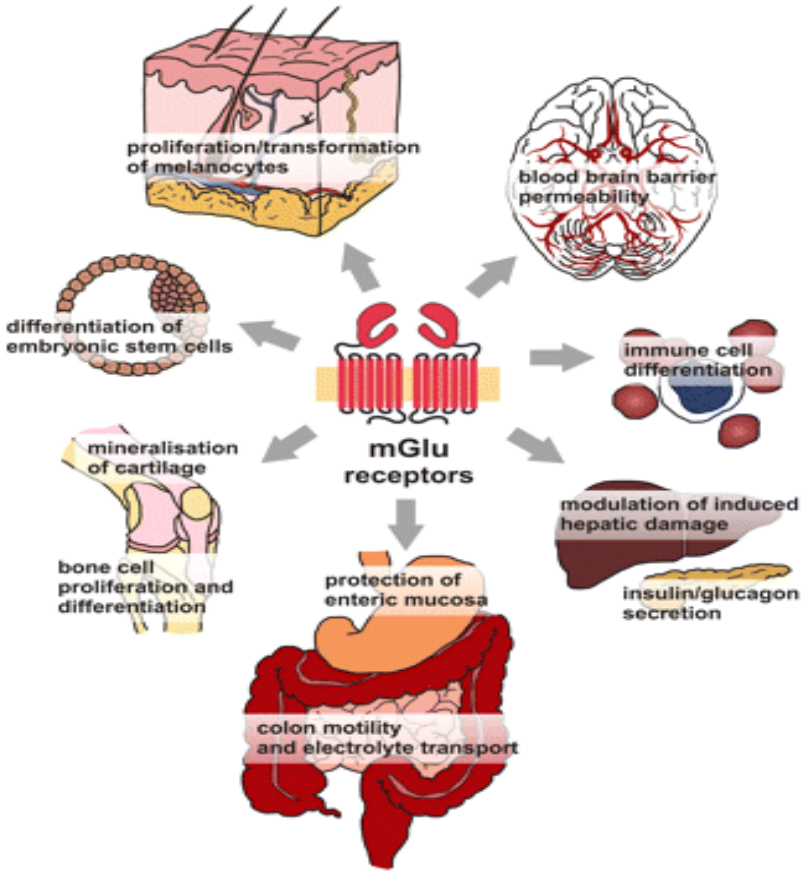

Figure 2.6: Summary of the roles of mGlu receptors in peripheral tissues [47].

mGluRs play important neuromodulatory roles throughout the brain as such they are targets for therapeutic intervention for a number of psychiatric and neurological disorders including anxiety disorders, depression, pain syndromes, epilepsy, Parkinson's disease and schizophrenia among others.

Chapter 3

Related Work

As introduced in the previous chapter, GPCRs are cell membrane proteins with a key role in many basic biological processes at the cellular level. The last decade has witnessed a fast evolution of their research, driven both by advances in genomics and systems biology, as well as by their increasingly clear importance in pharmacology.

Much of the functionality of a protein can be explained through its 3-D structure, which determines its ability for certain ligand binding. Despite intensive research, the 3-D structure is only fully determined for approximately a 12% of the human GPCR superfamily [31]. Alternatively, and under the assumption that the 3-D structure of the receptor is, at a basic label, fundamentally determined by the sequence of its constituent amino acids, the investigation of the functionality of a protein can be achieved through the analysis of such primary sequence when the information of the 3-D structure is not available. This is the case of class C GPCRs, the object of the current thesis, for which the first 3-D structure has only been determined in 2014.

Of late, research on sequence analysis has focused on the quantitative analysis of *aligned* sequence transformations as an alternative to the direct analysis of the raw sequence. Only recently, approaches based on machine learning and computational intelligence techniques have started to be used for the analysis of *alignment-free* sequence transformations. In this chapter, we summarily review recent research in this area related to this thesis, also developed as part of ongoing research project "*KAPPA AIM: Knowledge Acquisition in Pharmacoproteomics using Advanced Artificial Intelligence Methods*", funded by the Spanish MinECo. These works span different machine learning and computational intelligence approaches: from fully unsupervised, to semi-supervised and fully supervised.

In [44], alignment-free class C GPCR sequences were analyzed in a semi-supervised pattern recognition framework. That is, the models were trained with a limited

number of subtype-labeled sequences, with the goal of inferring the missing labels and thus, ultimately, classifying the available sequences. For this, latent variable models of the manifold learning family were employed, together with a manifold-graph building procedure. The experimental results indicated that the proposed semi-supervised models work best in situations of extreme class label scarcity, whereas semi-supervised Support Vector Machine (SVM)-based counterparts are competitive only when enough labeled sequences are available. The type of alignment-free GPCR transformation used was shown to have a key role in the classification accuracy results.

The work of Cárdenas and colleagues, instead, focuses on unsupervised approaches to reveal, through clustering and visualization, the natural grouping structure of the analyzed sequences. A key study focused on the visualization of misclassified class C CPCRs [55], using the GPCRDB database. The method for the visual exploration of misclassified sequences (from supervised classification methods) was based on a manifold learning model, namely the Generative Topographic Mapping (GTM), as well as on phylogenetic trees. It aimed to detect potential database labeling quality problems (in what could be understood as a label noise problem [64]). From the GTM visualization of class C GPCRs, it was observed that a reasonable level of subtype differentiation exists, but also that some subtypes, such as $GABA_B$, are more clearly separated from the rest than others, such as Pheromone or Veromonasal, which overlap to a high degree and are likely to be difficult to differentiate according to sequential primary information. It was also found that several mGluRs misclassified as Odorants, had a heterogeneous nature. Some were borderline cases likely to be sensitive to the choice of classifier, whereas others were clustered together in a position of the GTM visualization map that fully overlaps the most densely Odorant-population region. They were hypothesized that the latter were likely to be wrongly labeled in GPCRDB and, therefore, it was argued that expert data curators should re-examine these cases in the light of these results.

Cárdenas and co-workers, in a related study [56], used alignment-free sequence transformations for the exploratory visualization of mGluR subgroups using a kernelized variant of GTM, namely KGTM, using also the GPCRDB database. It was shown there that the visual representation of these data and, consequently, the type of knowledge that can be inferred from it, is at least partially dependent on the type of sequence transformation

employed. Moreover, the visualization provided only partial support of the three groups in which mGluRs are theoretically structured. *Group I* seems quite coherent in all representations, regardless data transformation type, whereas *Group II* is not clearly homogeneous according to any of the visualizations; similarly, limited homogeneity was observed in the four subtypes of mGluR that belong to *Group III* for all the transformations.

Related to [55], the study presented in [58] describes a systematic supervised approach to the analysis of class C GPCR misclassifications using SVMs for assisting the discovery of protein database labeling quality problems. From the experimental results, the existence of a number of instances that, independently of the sequence transformation method, induce classification errors was clearly observed. In the reported analyses, it was shown that the misclassifications of a sizable proportion of sequences were consistent and of a big enough magnitude as to prompt a recommendation for all these sequences to be revised by database curation experts due to the potential existence of label noise in the form of mislabeled class attributions.

A further related study [57] for the classification of class C GPCRs from alignment-free physicochemical transformations of their sequences, using SVMs, employed three sequence transformations: the amino acid composition transformation (AAC), the mean composition transformation (MC), and the auto cross covariance transformation (ACC). Moreover two measures were used to evaluate the test performance of the SVM multiclass trained classifier: the Accuracy and the Matthews correlation coefficient (MCC), which indicates how predictable the target variable is, knowing the other variables: its value ranges from -1 to 1 where 1 corresponds to a perfect classification, 0 to a random classification and -1 to complete misclassification. The best classification results were found for the ACC transformed dataset, achieving an accuracy of 93% and an MCC value of 0.91. It was thus shown that ACC transformed dataset has a clear advantage over the alternative transformations used and that SVMs are suitable for the analysis of this dataset. It was also shown that a classification *upper boundary* is likely to be reached in this problem, due to the previously mentioned noise label problems of the analyzed GPCRDB database.

Chapter 4

Materials and Methods

4.1 Materials 34

4.2 Methods 36

4.2.1 Alignment-Free Data Transformations 36

4.2.2 Data Clustering 38

4.2.2.1 Crisp and Fuzzy Partitions 39

4.2.3 K-means Clustering 42

4.2.4 Fuzzy c-Means Clustering 43

4.2.4.1 The Fuzzy c-Means Algorithm 44

4.2.4.2 Parameters of the FCM Algorithm 45

4.2.5 Assessment of stability in fuzzy clustering 48

This chapter includes, first, a description of the materials: the proteomics data in which the experimental analyses of this thesis are based. This is followed by a description of the methods applied in the data analysis: existing clustering methods, such as K-Means and Fuzzy C-Means; existing data transformation techniques for unaligned sequences; existing cluster stability assessment methods, such as the Separation and Concordance (SeCo) maps; but also a new extension of SeCo maps, of general purpose, which is proposed for fuzzy and probabilistic clustering methods.

4.1 Materials

The first GPCR crystal 3-D structure, that of rhodopsin, was determined in 2000 [33]. The number of researchers investigating the GPCRs 3-D structure has rapidly grown since

then. This is a difficult task though, as revealed by the fact that, currently, the 3-D structure of only 21 GPCRs has been fully determined, with the majority of them only in the last few years. By 2012, nine structures of class A GPCR had been published [34] , to which only two more were added in 2013 [35.36]. For the first time, in 2013, the 3-D structure of a GPCR not belonging to class A was determined. These were a structure belonging to the Frizzled class [37] and two more belonging to class B [38,39]. Finally, in 2014 the 3-D structure of an $mGluR_1$ receptor (class C) was determined [40]: the first class C GPCR 3-D structure to be unraveled.

This progress on the determination of GPCRs 3-D structure is made possible due to the either collaborative or competitive efforts from different laboratories around the world, following different approaches. In particular, one of the most active initiatives in the field is the GPCR Network, responsible for the determination of more than 50% of current crystal GPCR structures. The GPCR Network has an ongoing project, whose goal is to achieve 40%– 60% structural coverage of non-olfactory receptors for the period 2010 – 2015.

This thesis focuses on class C of GPCRs. As already mentioned, class C has become an increasingly important target for new therapies. The data set used for the purpose of our research was taken from GPCRDB [32]. This database-centered enterprise is a molecular-class information system that collects, combines, validates and stores large amounts of heterogeneous data related to GPCRs. GPCRDB divides the GPCR superfamily into five major classes as shown in Table 4.1, based on the ligand types, functions and sequence similarities. The data set analyzed was extracted from version 11.3.4 as of March 2011, of this database.

GPCR	Description
Class A	Rhodopsin-like
Class B	Secretin- like
Class C	Metabotropic glutamate/pheromone
Class D	Vomeronasal receptors (V1R and V3R)
Class E	Taste receptors T2R

Table 4.1: The five major classes of the GPCR superfamily according to GPCRDB.

The data set consists of 1,510 class C GPCRs sequences, which are further subdivided into 7 subtypes, including: 351 Metabotropic glutamate receptors (mGluR), 48 Calcium sensing receptors (CaSR), 208 $GABA_B$, 344 Vomeronasal (VN), 392 Pheromone (Ph), 102 Odorant (Od) and 65 Taste (Ta). The length of these sequences varies from 250 to 1995 amino acids (which clearly reveals the importance of alignment-free sequence transformation approaches).

In this thesis, the 351 available mGluR sequences were further investigated. These are in turn subdivided into eight recognized subtypes, namely *mGluR$_1$* to *mGluR$_8$* plus a group of mGluR-like sequences of unclear adscription. Specifically, there are 33 cases of $mGluR_1$, 26 of $mGluR_2$, 44 of $mGluR_3$, 23 of $mGluR_4$, 32 of $mGluR_5$, 15 of $mGluR_6$, 4 of $mGluR_7$, 98 of $mGluR_8$ and 76 cases of mGluR-like sequences. The main eight subtypes can also be grouped into three categories, as shown in Table 4.2, according to their protein coupling behavior. The final number of mGluR cases that were used is 256 out of 351. The 76 cases of mGluR-like were removed, as well as 19 cases with missing labels. Thus, the resulting number of cases for mGluR is 351 – 76 – 19 = 256 cases.

Class C groups	Members of group
Group I	$mGluR_1$ and $mGluR_5$
Group II	$mGluR_2$ and $mGluR_3$
Group III	$mGluR_4$, $mGluR_6$, $mGluR_7$ and $mGluR_8$

Table 4.2: The main eight subtypes of metabotropic glutamate receptors grouped into three categories.

4.2 Methods

4.2.1 Alignment-Free Data Transformations

The unaligned symbolic sequences are unsuitable for direct analysis, but there exist many different primary sequence transformation techniques that overcome these limitations. In this thesis, alignment-free full sequence was used to limit information loss. Specifically, three alignment-free transformations were used: the amino acid composition transformation, the auto cross covariance transformation and the digram transformation. Detailed descriptions of the transformation can be found below.

- **Amino Acid Composition Transformation:** This simple transformation reflects the amino acid composition (AAC) of the primary sequence. The frequencies of the 20 sequence-constituting amino acids are calculated for each sequence and, as a result, an *N x 20* data matrix is obtained, where *N* is the number of instances in the data set. Thus, in our case the data matrix obtained for analysis is 1,510 x 20 (therefore, of a not overtly high dimensionality).

- **Auto Cross Covariance Transformation:** The ACC transformation is a more sophisticated one, and aims to capture the correlation of the physico-chemical amino acid descriptors along the sequence. The method relies on a multivariate approach where the primary amino acid sequences are translated into vectors based on the principal physicochemical properties of the amino acids, transforming the data into a uniform matrix by applying a modified autocross-covariance transform [41]. First, the physico-chemical properties are represented by means of the five z-scores of each amino-acid as described in [42]. Then the Auto Covariance (AC) and Cross Covariance (CC) variables are computed on this first transformation. These variables measure respectively the correlation of the same descriptor (AC) or 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 up to a maximum lag, *l*. This transformation generates an *N* x ($z^2 \cdot l$) matrix, where z = 5 is the number of descriptors. The maximal lag that was used for the ACC transformation is l = 13, which was found in previous studies to provide the best accuracy for this dataset in [44]. Thus, the matrix is *N* x *325*, where *N* = 1,510 (thus, the dimensionality of the data is moderately high, with a lower than 5 to 1 ratio of data sequences to variables).

- **Digram Transformation:** The digram transformation is a particular instance of the more general n-gram transformation. It considers the frequencies of occurrence of any given pair of AAs. The *n*-gram concept has previously been used in protein analysis [43]. This particular transformation generated an *N* x *400* matrix, where *N* = 1,510 (complete class C data set).

4.2.2 Data clustering

For the purpose of class C GPCR subtyping, both crisp and fuzzy clustering techniques were used, namely K-means and its fuzzy variant, Fuzzy c-Means (FCM). A more detailed description of crispy and fuzzy clustering, as well as of the K-means and FCM algorithms, is provided next.

Data clustering is the process of dividing data elements into clusters so that items in the same cluster are as similar as possible, and items in different clusters are as dissimilar as possible. Clustering techniques are mostly unsupervised methods, that is, no group labels are used to generate the data model (or, in different words, train the algorithm). Depending on the nature of the data and the purpose for which clustering is being used, different measures of similarity may be used to assign data items into clusters, where the similarity measure controls how the clusters are formed. Some examples of measures that can be used as in clustering include distance, connectivity, and intensity.

Data can reveal clusters of different geometrical shapes, sizes and densities as illustrated in Figure 4.1. While clusters (a) are spherical, clusters (b) to (d) can be characterized as linear and nonlinear subspaces of the data space. The performance of most clustering algorithms is influenced not only by the geometrical shapes and densities of the individual clusters, but also by the spatial relations and distances among the clusters. Clusters can be well-separated, continuously connected to each other, or overlapping each other.

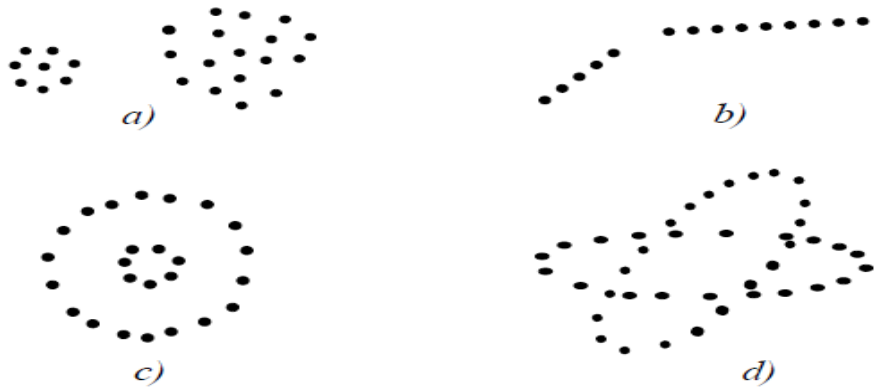

Figure 4.1: Illustration of clusters of different shapes and dimensions.

Many clustering algorithms have been introduced in the literature. Since clusters can formally be seen as subsets of the data set, one possible classification of clustering methods can be according to whether the subsets are fuzzy or crisp.

In crisp clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. Fuzzy clustering methods, also referred to as soft clustering, allow the objects to belong to several clusters simultaneously, with different degrees of membership (between 0 and 1). These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

4.2.2.1 Crisp and Fuzzy Partitions

Clustering techniques can be applied to data that are quantitative (numerical), qualitative (categorical), or a mixture of both. The data used in this thesis are of quantitative, real-valued nature. The data are typically observations of some physical process. Each observation consists of n measured variables, called features or attributes, grouped into an n-dimensional column vector. The data can be represented as an n x N matrix, where each column is an observation and each row represent a feature or attribute. The objective of a clustering algorithm is to partition a dataset Z into c clusters. The number of clusters c is a variable of the model.

Crisp Partition:

Using classical sets, a crisp partition of Z can be defined as a family of subsets as shown in equation 4.1 with the following properties (4.2a, 4.2b, 4.2c) [45]:

$$\{A_i \mid 1 \leq i \leq c\} \subset \mathcal{P}(\mathbf{Z}) \qquad 4.1$$

$$\bigcup_{i=1}^{c} A_i = \mathbf{Z}, \qquad 4.2a$$

$$A_i \cap A_j = \emptyset, \quad 1 \leq i \neq j \leq c, \qquad 4.2b$$

$$\emptyset \subset A_i \subset \mathbf{Z}, \quad 1 \leq i \leq c\,. \qquad 4.2c$$

The union of the subsets A_i contains all the data, as described in equation 4.2a. The subsets must be disjoint, as stated by equation 4.2b, and none of them is empty nor contains all the data in Z (4.2c). The partition of the data into clusters is conveniently represented by the partition matrix $\mathbf{U} = [\,\mu_{ik}\,]_{cxN.}$ The i^{th} row of this matrix contains values of the membership function μ_i of the i^{th} subset A_i of Z. It follows, from the above equations (4.2), that the elements of $\mathbf{U}$ must satisfy the following conditions:

$$\mu_{ik} \in \{0, 1\}, \quad 1 \leq i \leq c, \quad 1 \leq k \leq N, \qquad 4.3a$$

$$\sum_{i=1}^{c} \mu_{ik} = 1, \quad 1 \leq k \leq N, \qquad 4.3b$$

$$0 < \sum_{k=1}^{N} \mu_{ik} < N, \quad 1 \leq i \leq c\,. \qquad 4.3c$$

The space of all possible crisp partition matrices for Z, called the crisp partitioning space [45] is thus defined by

$$M_{hc} = \left\{ \mathbf{U} \in \mathbb{R}^{c \times N} \middle| \mu_{ik} \in \{0, 1\}, \forall i, k;\ \sum_{i=1}^{c} \mu_{ik} = 1, \forall k;\ 0 < \sum_{k=1}^{N} \mu_{ik} < N, \forall i \right\}$$

Fuzzy Partition:

Generalization of the crisp partition to the fuzzy case follows directly by allowing μ to obtain real values in [0, 1]. Conditions for a fuzzy partition matrix, analogous to (4.3) are

given by:

$$\mu_{ik} \in [0,1], \quad 1 \leq i \leq c, \quad 1 \leq k \leq N, \qquad \text{4.4a}$$

$$\sum_{i=1}^{c} \mu_{ik} = 1, \quad 1 \leq k \leq N, \qquad \text{4.4b}$$

$$0 < \sum_{i=1}^{N} \mu_{ik} < N, \quad 1 \leq i \leq c\,. \qquad \text{4.4c}$$

The i^{th} row of the fuzzy partition matrix $\mathbf{U}$ contains values of the i^{th} membership function of the fuzzy subset A_i of Z. The sum of each column must be equal to 1, as described by the 4.4b equation. Thus the total membership of each observation, z_k, in Z equals one. The fuzzy partitioning space for Z is the set:

$$M_{fc} = \left\{ \mathbf{U} \in \mathbb{R}^{c \times N} \middle| \mu_{ik} \in [0,1], \forall i,k;\ \sum_{i=1}^{c} \mu_{ik} = 1, \forall k;\ 0 < \sum_{k=1}^{N} \mu_{ik} < N, \forall i \right\}$$

Example: Let's assume that we have a data set $Z = \{ z_1, \ldots, z_{10} \}$ as shown in Figure 4.2.

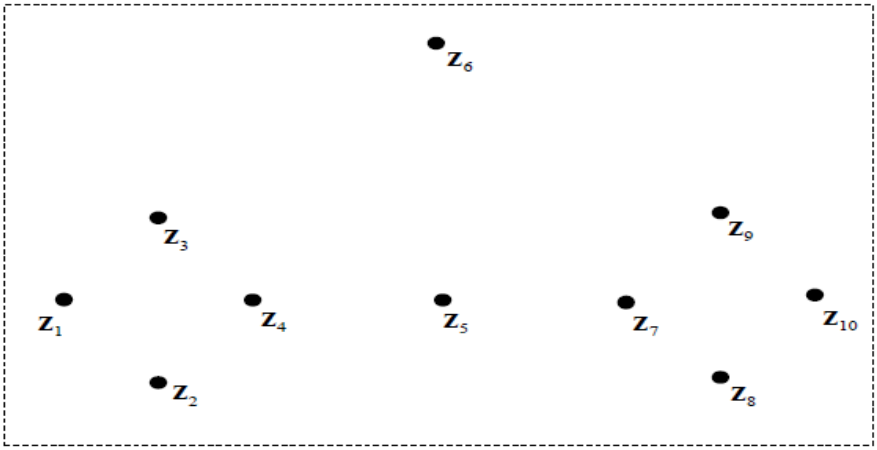

Figure 4.2: Representation of data set Z in a two dimensional space.

One possible hard partition (out of 2^{10} possible hard partitions), $\mathbf{U}$, of the data set Z, into two clusters is:

$$\mathbf{U} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

The first row of $\mathbf{U}$ defines point-wise the characteristic function for the first subset of Z, A_1, and the second row defines the characteristic function of the second subset of Z, A_2. Each sample must be assigned exclusively to one subset (cluster) of the partition. In this case, both the boundary point $\mathbf{z}_5$ and the outlier $\mathbf{z}_6$ have been assigned to A_1. It is clear that a hard partitioning may not give a realistic picture of the underlying data. Boundary data points may represent patterns with a mixture of properties of data in A_1 and A_2, and therefore cannot be fully assigned to either of these classes, or do they constitute a separate class.

Now if we are using a fuzzy clustering algorithm, one of the infinitely many possible fuzzy partitions in Z that we may have is:

$$\mathbf{U} = \begin{bmatrix} 1.0 & 1.0 & 1.0 & 0.8 & 0.5 & 0.5 & 0.2 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.2 & 0.5 & 0.5 & 0.8 & 1.0 & 1.0 & 1.0 \end{bmatrix}$$

The boundary point $\mathbf{z}_5$ has now a membership degree of 0.5 in both classes, which correctly reflects its position in the middle between the two clusters. Note, however, that the outlier $\mathbf{z}_6$ has the same pair of membership degrees, even though it is further from the two clusters, and thus can be considered less typical of both A_1 and A_2 than $\mathbf{z}_5$. This is because condition (4.4b) requires that the sum of memberships of each point equals one.

4.2.3 K-means Clustering

K-means clustering is an unsupervised learning algorithm that aims to partition n data points into a certain number of clusters, k, fixed priori, so as to minimize the within cluster sum of squares. In other words, its goal is to minimize the following objective function:

$$\sum_{i=1}^{K} \sum_{\mathbf{x} \in \Gamma_i} \|\mathbf{x} - \mu_i\|^2 \qquad (4.5)$$

where: K: is the number of clusters;

x: is a data point belonging to cluster Γ_i, and

μ_i : is the mean of the points in cluster Γ_i.

The K-means algorithm consists on the following steps:

1. Choosing the value of k.
2. Selecting k random instances $\{\mathbf{s}_1, \ldots, \mathbf{s}_k\}$ to be the initial cluster centers.
3. Calculating the distance between each data point and each cluster center.
4. Assigning class membership (to which cluster each data point belongs).
   That is, for each data point $\mathbf{x}_i$, assign $\mathbf{x}_i$ to the cluster c_j such that $\mathrm{d}(\mathbf{x}_i, c_j)$ is minimal.
5. Update the cluster centers:
   For each cluster c_j
   $\mathbf{s}_j = \mu(c_j)$ $\qquad \vec{\mu}(c) = \frac{1}{|c|}\sum_{\vec{x}\in c}\vec{x}$
6. Stopping criterion:
   If none of the data points changed membership, exit;
   otherwise, go to step 3

4.2.4 Fuzzy c-Means Clustering

FCM is a method of clustering that allows data points to belong to one or more clusters. Most analytical fuzzy clustering algorithms are based on optimization of the basic c-means objective function, or some modification of it. The objective function of FCM is:

$$J(\mathbf{Z};\mathbf{U},\mathbf{V}) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m \|\mathbf{z}_k - \mathbf{v}_i\|_{\mathbf{A}}^2 \qquad (4.6)$$

where:

- $\mathbf{U} = [\mu_{ik}] \in M_{fc}$

is a real-valued c x N matrix, that represents the fuzzy partition matrix of $\mathbf{Z}$: c is the number of clusters and N is the number of observation in the data.

- $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c], \quad \mathbf{v}_i \in \mathbb{R}^n$

  is a vector of cluster centers, which have to be determined.

- $D^2_{ik\mathbf{A}} = \|\mathbf{z}_k - \mathbf{v}_i\|^2_{\mathbf{A}} = (\mathbf{z}_k - \mathbf{v}_i)^T \mathbf{A}(\mathbf{z}_k - \mathbf{v}_i)$

  is a squared inner-product distance norm, and

- $m \in [1, \infty)$

  is a parameter which determines the fuzziness of the resulting clusters. The value of the objective function can be seen as measure of the total variance of $\mathbf{z}_k$ from $\mathbf{v}_i$.

4.2.4.1 The Fuzzy c-Means Algorithm

The objective of the FCM algorithm is the minimization of the objective function. This minimization represents a nonlinear optimization problem that can be solved by using a variety of methods, including iterative minimization, simulated annealing or genetic algorithms.

FCM Algorithm Steps:

Given a data set $\mathbf{Z}$, choose the number of clusters $1 < c < N$, the weighting exponent $m > 1$, the termination tolerance $\varepsilon > 0$ and the norm-inducing matrix $\mathbf{A}$. Initialize the cluster center matrix $\mathrm{V}^{(0)}$ randomly.

Repeat for l=1,2,...

Step 1: Compute the distance between each data point and each cluster center

$$D^2_{ik\mathbf{A}} = (\mathbf{z}_k - \mathbf{v}_i^{(l)})^T \mathbf{A}(\mathbf{z}_k - \mathbf{v}_i^{(l)}), \quad 1 \le i \le c, \quad 1 \le k \le N$$

Step 2: Update the fuzzy partition matrix $\mathbf{U}$. Compute the membership values of

each data point to each cluster

for $1 \leq k \leq N$

if $D_{ik\mathbf{A}} > 0$ for all $i = 1, 2, \ldots, c$

$$\mu_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{c} (D_{ik\mathbf{A}} / D_{jk\mathbf{A}})^{2/(m-1)}},$$

otherwise

$$\mu_{ik}^{(l)} = 0 \text{ if } D_{ik\mathbf{A}} > 0, \text{ and } \mu_{ik}^{(l)} \in [0, 1] \text{ with } \sum_{i=1}^{c} \mu_{ik}^{(l)} = 1$$

Step 3: Compute the objective function value

$$\textbf{obj}^{(l)} = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik}^{(l)})^m \, D_{ik\mathbf{A}}^2$$

Step 4: Compute the new cluster centers and update the cluster centers matrix

$$\mathbf{v}_i^{(l)} = \frac{\sum_{k=1}^{N} \left(\mu_{ik}^{(l-1)}\right)^m \mathbf{z}_k}{\sum_{k=1}^{N} \left(\mu_{ik}^{(l-1)}\right)^m}, \quad 1 \leq i \leq c$$

until $\|\text{obj}^{(l)} - \text{obj}^{(I-1)}\| < \varepsilon$

When the algorithm is terminated, the matrix **U** contains the final fuzzy partitions for each data point. As previously mentioned, **U** contains the membership degree, [0,1] to which a data point belongs to a specific cluster. The most common method to assign the data points to specific clusters entails finding the maximum membership value for each data point and assigning the point to the corresponding cluster.

4.2.4.2 Parameters of the FCM Algorithm

Before using the FCM algorithm, the following parameters must be set: the number of clusters, c, the 'fuzziness' exponent, m, the termination tolerance, ε, and the norm-inducing matrix, $\mathbf{A}$.

Number of clusters: The number of clusters c is the most important parameter, in the sense that the remaining parameters have less influence on the resulting partition. When clustering real data without any a priori information about the structures in the data, one usually has to make assumptions about the number of underlying clusters. The chosen clustering algorithm then searches for c clusters, regardless of whether they are really present in the data or not. Two main approaches to determining the appropriate number of clusters in data can be distinguished:

1. *Validity measures*. Validity measures are scalar indices that assess the goodness of the obtained partition. Clustering algorithms generally aim at locating well separated and compact clusters. When the number of clusters is chosen equal to the number of groups that actually exist in the data, it can be expected that the clustering algorithm will identify them correctly. When this is not the case, misclassifications appear, and the clusters are not likely to be well separated and compact. Hence, most cluster validity measures are designed to quantify the separation and the compactness of the clusters. However, as Bezdek points out in his paper [45], the concept of cluster validity is open to interpretation and can be formulated in different ways. Consequently, many validity measures have been introduced in the literature. For the FCM algorithm, the Xie-Beni index [62]:

$$\chi(\mathbf{Z};\mathbf{U},\mathbf{V}) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{N}\mu_{ik}^{m} \parallel \mathbf{z}_k - \mathbf{v}_i \parallel^2}{c \cdot \min_{i\neq j}\left(\parallel \mathbf{v}_i - \mathbf{v}_j \parallel^2\right)}$$

is an example that has been found to perform well in practice. This index can be interpreted as the ratio of the total within-group variance and the separation of the

cluster centers.

The best partition minimizes the value of $\chi(\mathbf{Z};\mathbf{U},\mathbf{V})$.

2. *Iterative merging or insertion of clusters*. The basic idea of cluster merging is to start with a sufficiently large number of clusters, and successively reduce this number by merging clusters that are similar (compatible) with respect to some well-defined criteria. One can also adopt an opposite approach, i.e., start with a small number of clusters and iteratively insert clusters in the regions where the data points have low degree of membership for the existing clusters.

Fuzziness Parameter: The weighting exponent *m* is a rather important parameter as well, because it significantly influences the fuzziness of the resulting partition. As *m* approaches one from above, the partition becomes hard ($\mu_{ik} \in \{0, 1\}$) and v_i are ordinary means of the clusters. As $m \rightarrow \infty$, the partition becomes completely fuzzy ($\mu_{ik} = 1/c$) and the cluster means are all equal to the mean of Z. Usually, $m = 2$ is initially chosen.

Termination Criterion: The FCM algorithm stops iterating when the norm of the difference between the values of the objective function in two successive iterations is smaller than the termination parameter ε. The usual choice is $\varepsilon = 0.001$, even though $\varepsilon = 0.01$ works well in most cases, while drastically reducing the computing times.

Norm-Inducing Matrix: The shape of the clusters is determined by the choice of the matrix $\mathbf{A}$ in the distance measure as shown in the distance equation in section 4.2.2.2. A common choice is $\mathbf{A} = \mathbf{I}$, which gives the standard Euclidean norm:

$$D_{ik}^2 = (\mathbf{z}_k - \mathbf{v}_i)^T (\mathbf{z}_k - \mathbf{v}_i).$$

Another choice for $\mathbf{A}$ is a diagonal matrix that accounts for different variances in the directions of the coordinate axes of $\mathbf{Z}$:

$$\mathbf{A} = \begin{bmatrix} (1/\sigma_1)^2 & 0 & \cdots & 0 \\ 0 & (1/\sigma_2)^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (1/\sigma_n)^2 \end{bmatrix}$$

This matrix induces a diagonal norm with *n*-dimensions. Finally, $\mathbf{A}$ can be defined as the inverse of the covariance matrix of $\mathbf{Z}$: $\mathbf{A} = \mathbf{R}^{-1}$, with

$$\mathbf{R} = \frac{1}{N} \sum_{k=1}^{N} (\mathbf{z}_k - \bar{\mathbf{z}})(\mathbf{z}_k - \bar{\mathbf{z}})^T$$

In this case, $\mathbf{A}$ induces the Mahalanobis norm. The norm influences the clustering criterion by changing the measure of dissimilarity. The Euclidean norm induces hyperspherical clusters. Both the diagonal and the Mahalanobis norm generate hyperellipsoidal clusters. With the diagonal norm, the axes of the hyperellipsoids are parallel to the coordinate axes, while with the Mahalanobis norm the orientation of the hyperellipsoid is arbitrary, as shown in Figure 4.3.

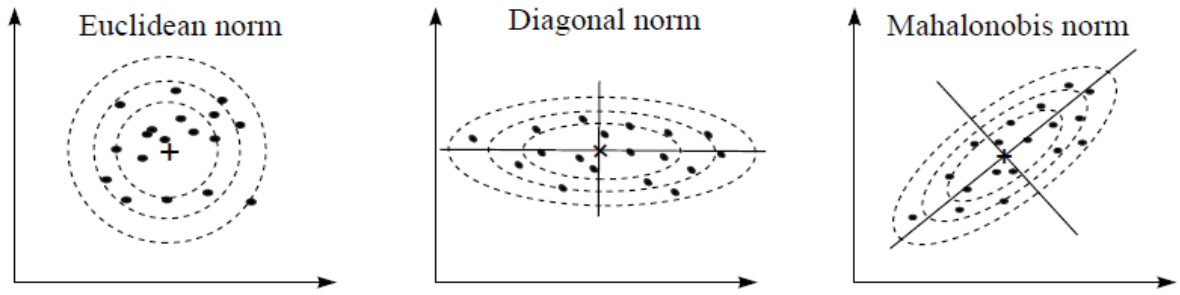

Figure 4.3: Different distance norms used in fuzzy clustering

4.2.5 Assessment of stability in fuzzy clustering

After decades of intensive use, K-means is still a common choice for crisp clustering in real-world applications [63]. As previously mentioned, variants of the algorithm, such as

FCM, have extended the model to account for partial degrees of membership. K-means limitations have been well-studied over the years and include the lack of a closed criterion for the choice of the number of clusters *K* and the fact that, under different initializations, the algorithm may yield very different solutions. K-means objective is finding the set of cluster centroids, Γ, that minimizes a SSQ error in the form of Eq. 4.5. FCM generalizes this objective function to become Eq. 4.6. Even if finding a minimum error is a central objective of K-means and FCM, the stability of the clustering solution is also a relevant one. Solutions that are reproducible are often required in practical clustering applications.

Recent experimental evidence [65] has shown that K-means solutions that might be expected to be similar according to the final value of the objective function may in fact be quite dissimilar, and that this effect increases with the value of *K.* This suggests the convenience of using the objective function as a criterion of model optimality *only in combination with some cluster stability criterion* if we aim to achieve cluster partition reproducibility. One such combined approach is based on the use of the Separation / Concordance (SeCo) maps [65], which are summarily described next. In this chapter, this criterion is generalized to cover also fuzzy and probabilistic clustering techniques.

Assuming that solutions that strike a balance between low error and high stability ought to be sought, Lisboa *et al.* [65] proposed a clustering solution analysis framework based on the calculation of SeCo maps for settings in which multiple random initializations of K-means for different values of *K* were used. SeCo maps display, in corresponding coordinate axes, the ΔSSQ, which is calculated as the total SSQ minus the within-cluster SSQ:

$$\sum_{i=1}^{K}\sum_{j=1}^{N}\|\mathbf{x_j}-\mu_i\|^2-\sum_{i=1}^{K}\sum_{\mathbf{x_j}\in\Gamma_i}\|\mathbf{x_j}-\mu_i\|^2 \qquad (4.7)$$

against a concordance index (CI) quantifying the stability. The CI is calculated as the median of the (nin -1) pairwise Cramér's V index calculations for nin initializations. Cramér's V stability index is a variation of Pearson's chi-squared statistic (χ^2). For two different cluster partitions of the same data into, in turn, K and K' clusters, Cramér's V

index is calculated as:

$$C_v = \sqrt{\frac{\chi^2}{N.\min(K-1, K'-1)}}$$

where,

$$\chi^2 = \sum_{k=1}^{K}\sum_{k'=1}^{K'}(O_{kk'} - E_{kk'})^2 / E_{kk'} \qquad (4.8)$$

Here, $\mathbf{O}$ is an observed contingency table ($K \times K'$) matrix, whose values $O_{kk'}$ indicate the number of instances in $\mathbf{X}$ that have been assigned to cluster k in one run of the algorithm and to cluster k' in another run. The $K \times K'$ matrix $\mathbf{E}$ contains the corresponding expected values for independent cluster allocations, calculated as:

$$E_{kk'} = \frac{1}{N}\left(\sum_{j=1}^{K'} O_{kj} \sum_{i=1}^{K} O_{ik'}\right) \qquad (4.9)$$

This use of contingency tables for the calculation of Cramér's V index as the basis for the CI used in the SeCo approach is suitable for crisp cluster assignments such as those provided by K-Means. For soft assignments such as those provided by FCM or probabilistic techniques such as Gaussian Mixture Models, instead, such contingency tables occlude the richness of the cluster solution by requiring the assignment of instances to clusters to be based on the highest degree of membership or probability.

In this thesis, we propose a variation of contingency tables that better suits the characteristics of fuzzy and probabilistic models. Elements in what we call weighted observed ($\omega\mathbf{O}$) contingency tables will now be calculated, following the notation of Eq.4.6, as:

$$wO_{kk'} = sum_{i=1}^{N} \mu_{ik}\,\mu_{ik'} \qquad (4.10)$$

where, for FCM this is, for data instance i, the product of the degree of membership to

cluster k in a first run of the algorithm and the degree of membership to cluster k' in a second run. Consequently, we can obtain a weighted expected ($\omega\mathbf{E}$) contingency table matrix whose elements are defined as:

$$wE_{kk'} = \frac{(\sum_{j=1}^{K} wO_{kj} \sum_{i=1}^{K'} wO_{ik'})}{N} \qquad (4.11)$$

This leads to the definition of a new weighted Cramér's V index, where $\mathbf{O}$ is replaced by $\omega\mathbf{O}$ and $\mathbf{E}$ by $\omega\mathbf{E}$ in the calculation of Eq.4.8.

If FCM estimated that all instances had a degree of membership of 1 for a single cluster, the weighted Cramér's V index would reduce to its standard formulation. This is unlikely to happen, which means that the proposed index will lead to lower levels of CI in SeCo. This should, therefore, be not only a conservative concordance estimator, but also a more reliable clustering assessment tool, capable of distinguishing solutions with varying levels of certainty.

Note that SeCo can be used as a flexible informative tool for the choice of adequate values of the *K* parameter (number of clusters). This is equally true when using the modified index, but, in this case, there should be no bias in favor of “over-optimistic” solutions.

Chapter 5

Experimental Study

5.1 Experiments with a fixed number of clusters 54

5.1.1 Fuzzy c-Means for class C GPCRs 55

5.1.2 K-means for class C GPCRs 63

5.2 Experiments with varying number of clusters: Cluster Stability Analysis 67

5.2.1 Full Separation Concordance (SeCo) map 69

5.2.1.1 Class C GPCR results 69

5.2.1.1 mgluR results 73

5.2.2 Thresholding the objective function 76

5.2.2.1 Class C GPCR results 77

5.2.2.2 mGluR results 80

The discrimination of GPCRs into types and subtypes according to the amino acid symbolic sequences that describe them can provide useful insights for the design of targeted pharmacological drugs.

The data set used for the purpose of our research, as explained in the previous chapter, contains 1,510 class C GPCR sequences, the length of which varies from 250 to 1995 amino acids. Thus, the unaligned symbolic sequences are mostly unsuitable for direct analysis. For this reason, the alignment-free sequence transformations described in chapter 4 (AAC, ACC and Digram) were used to transform the data into a format amenable to be modeled by the K-means and FCM algorithms.

Experiments were carried out both using the complete class C GPCR dataset,

including the 1,510 available amino acid sequences, and the mGluR class C subtype, including the available 256 amino acid sequences. The six resulting data sets (2 data sources x 3 data transformations) were used for analysis, as summarized in Table 5.1.

GPCR subtype	Sequence Transformation	Name	Data Matrix Dimensions
Class C GPCRs	AAC	CGPCR_AAC	1510x20
Class C GPCRs	ACC	CGPCR_ACC	1510x325
Class C GPCRs	Digram	CGPCR_Digram	1510x400
mGluR	AAC	mGluR _AAC	256x20
mGluR	ACC	mGluR _ACC	256x325
mGluR	Digram	mGluR _Digram	256x400

Table 5.1: List of the six datasets used in these experiments. The three sequence transformations for class C GPCRs, plus the three sequence transformations for mGluRs.

FCM was compared to the standard K-means algorithm in two different experimental settings:

The first one assumes fixed number of clusters and that this is the same as the number of known receptor subtypes defined in GPCRDB. The rationale behind this part of the experiments was to investigate to what extent the obtained clusters naturally resembled the *standard* existing subtype definition when data are transformed the way we have, and how well each algorithm separates the data. According to this rationale, the achieved accuracy of the K-means and FCM algorithms was not our main concern, given that we expected many of the subtypes to overlap to different degrees. Instead, our goal was to extract knowledge about the levels of subtype (class) specificity in each cluster and to ascertain to what extent the subtypes of each data set were well separated.

The second setting removes the constraint of fixing the number of clusters *a priori* and focuses on the analysis of the stability of the clustering results, applying a methodology that allows a trade-off between low clustering error and high clustering stability and, as a result, reproducible solutions that are applicable in real problems.

The experimental results of the first set of experiments are presented and discussed in section 5.1 and those of the second set of experiments are presented and discussed in section 5.2.

All algorithms and experimental scripts were implemented and tested using Matlab in version R2012b. Some experiments were run using the computation cluster at the Computer Science Department of the *Universitat Politècnica de Catalunya* (UPC), managed by the rdlab[1].

5.1 Experiments with a fixed number of clusters

The six datasets listed in Table 5.1 were fed to the FCM and K-means algorithms to explore in what extent the obtained clusters resembled the *standard* subtype definition. More in detail, the class (subtype) specificity for each cluster for each dataset was measured and the results are provided in the following sub-sections along with class-entropy measures. This will inform us to what extent the clusters extracted by K-means and FCM algorithms correspond (or not) to the theoretically labeled subtypes. The class-entropy for a given cluster *k* will take the general form:

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

where *j* is one of the C = 7 class C GPCR subtypes, or C = 3 mGluR subtypes, and $p_{kj} = m_{kj} / m_k$, where, in turn, m_k is the number of sequences in cluster k and m_{kj} is the number of subtype *j* sequences in cluster k. The entropy of a pure cluster (with sequences of a single subtype assigned) will thus be zero because $S_k = ln(1) = 0$, whereas a maximum entropy will be reached when all subtypes are equally represented in a cluster.

[1] http://rdlab.cs.upc.edu

A discussion of the results is also included at the end of this section there is a comparison of the two algorithms based on the results.

5.1.1 Fuzzy c-Means for class C GPCRs

As mentioned and in the previous chapter, some specific parameters have to be specified in the setting of the FCM algorithm. The initial parameters specified for each one of them are described below.

Number of clusters, *c:* For the class C GPCRs datasets the number of clusters c was specified as 7 (the same as the number of standard subtypes defined in GPCRDB) and for the mGluR datasets the number of clusters c was specified as 3 (in this case to match the number of main groups of mGluR, as described in Table 4.2).

Fuzziness Parameter, *m:*

A major problem in applying the FCM method for clustering amino acid sequence data is the choice of the fuzziness parameter. The minimization of the objective function depends on the norm metric for distance calculation and on the choice of m. In existing literature, the most common choice for m is a value of 2. However, our preliminary experiments found that this choice did not yield reasonable results for our data sets.

As stated in [54], the commonly used value $m = 2$ may not be appropriate for some data sets, and that optimal value for m vary widely from one data set to another. In this paper, authors proposed a method for finding an appropriate value for m.

As it was shown in [45], when m tends to infinity, values of the $\mathbf{U}$ partition matrix (u_{ci}) converge to $1/c$ (where c is the number of clusters). Thus, for a given data set, there is an upper bound value for m, m_{ub}, above which the membership values resulting from FCM are equal to $1/c$. Thus, the upper bound of m for the GPCR data set is between 1 and 2. The suggested method is as follows: for each chosen value of m between 1 and 2, the coefficient of variation cv of the set of distances between

$$Y_m = \{ [d^2(z_i, z_c)]^{1/m-1} ; c \neq i = 1,2,..., N \}$$

coefficient variation equation:

$$cv\{Y_m\} = \frac{\sigma_{Y_m}}{\bar{Y}_m} \approx 0.03p$$

where:

σ_{Y_m} : is the standard deviation,

$\bar{Y}_m$: is the mean of the set Y_m

P : is the dimensionality of the data (number of attributes)

After the determination of the upper bound, m_{ub}, the determination of m must be done. In [54], a value of $m = 1 + m_0$ was chosen, where $m_0 = 1$ if $m_{ub} \geq 10$ and $m_0 = m_{ub}/10$ if $m_{ub} \leq 10$.

Table 5.2 displays the upper bound of m found for each dataset. As observed from several experiments, the value of m chosen from the first data set from the above equations was not the optimal one. For this reason, the accuracy obtained from the FCM algorithm with every value of m in the range of [1.1, m_{ub}] was calculated for this data set and the m that yielded the best accuracy was chosen. The values of m chosen for each dataset are displayed in Table 5.2.

Data Set	N	p	c	m_{ub}	m
CGPCR_AAC	1510	20	7	1.50	1.30
CGPCR_ACC	1510	325	7	1.10	1.10
CGPCR_Digram	1510	400	7	1.10	1.10
mGluR _AAC	256	20	3	1.90	1.19
mGluR _ACC	256	325	3	1.95	1.19
mGluR _Digram	256	400	3	1.90	1.19

Table 5.2: Fuzziness Parameter *m* for each data set. In this table, the upper bound of the fuzziness parameter *m* for each data set is displayed, as well as the chosen value of *m* for each one of them. *N* is the number of cases in the data set; *p* the number of attributes; and *c* the number of clusters.

Termination Criterion, *ε:*

The termination criterion, ε, was set, for all data sets, at a value of 1e-6.

Norm-Inducing Matrix, *A:*

The norm-inducing matrix that was taken to be $\mathbf{A} = \mathbf{I}$, which results in the standard Euclidean norm, for all the data sets.

Maximum Number of Iterations, *maxIter:*

The maximum number of iterations, *maxIter*, was chosen to be 500. In all cases, the algorithm was terminated before reaching this threshold value, according to the termination criterion.

After setting all initial parameters, the FCM algorithm was applied to the data sets. On termination of the algorithm, the final fuzzy partition of the data was made available

(matrix **U**) and stored. Based on the membership values of **U**, each data point was assigned to the cluster that had the highest membership value (crisp assignment) and the class (subtype) specificity of each cluster for each data transformation was measured as reported in Figs. 5.1, 5.2 and 5.3. Subtype- (class-)entropies were also calculated as reported in Tables 5.3, 5.4 and 5.5.

Class C GPCR results

- **AAC Transformation**

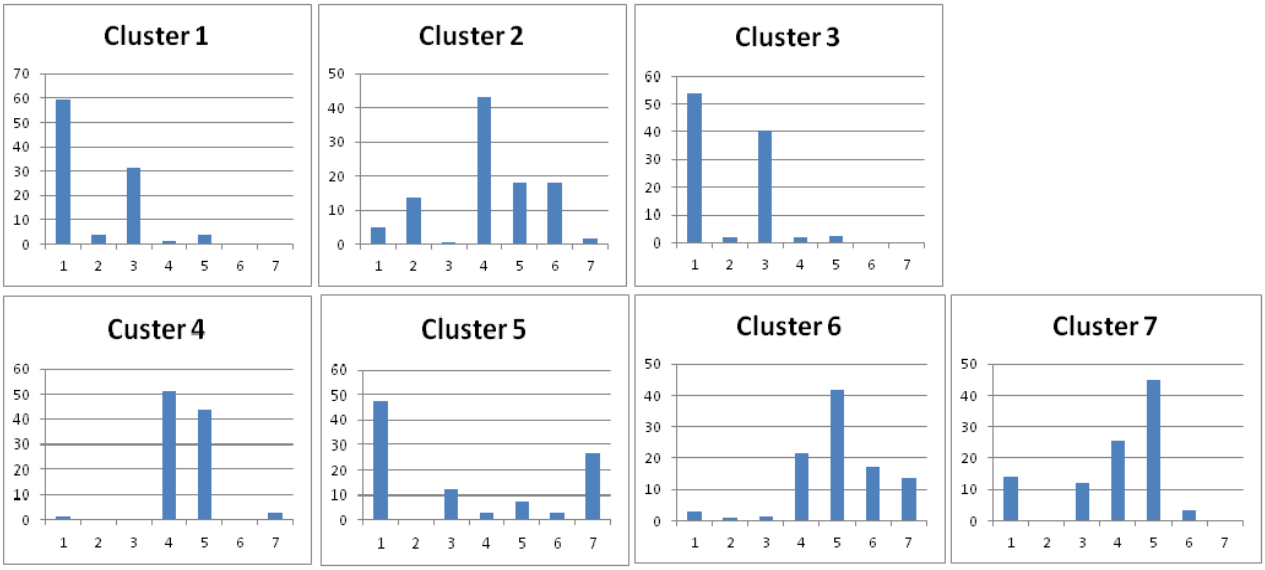

Figure 5.1: Class specificity for each cluster of the complete C GPCR data set with the AAC transformation.

Cluster	# of instances	Entropy
Cluster 1	245	1.45
Cluster 2	239	2.16
Cluster 3	200	1.34
Cluster 4	193	1.30
Cluster 5	67	1.97
Cluster 6	263	2.15
Cluster 7	303	1.95
Total Entropy		**1.77**

Table 5.3: : Entropy measure for the complete C GPCR data set with the AAC transformation.

- **ACC Transformation**

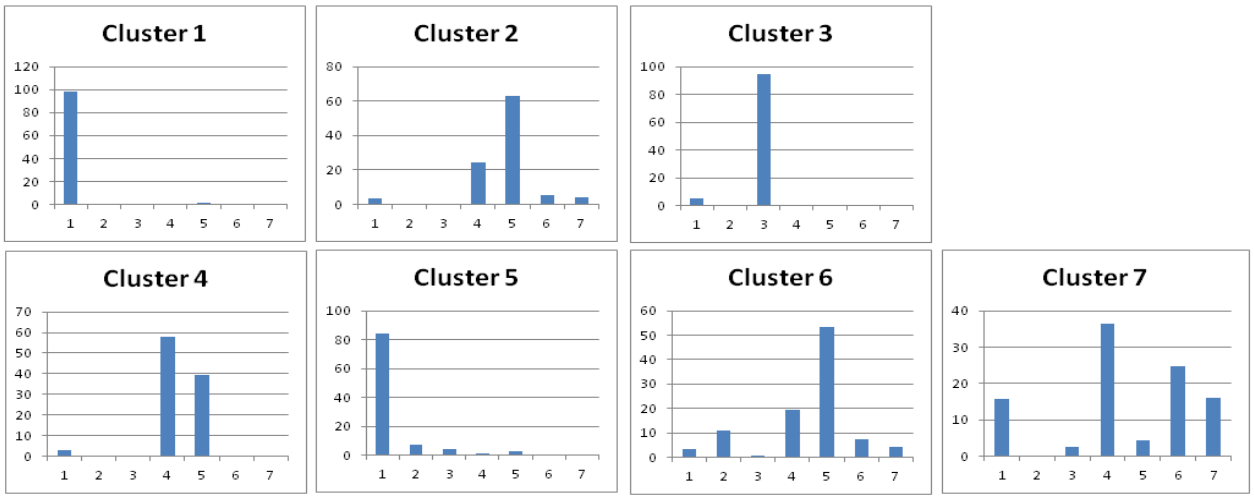

Figure 5.2: Class specificity for each cluster of the complete C GPCR data set with the ACC transformation.

Cluster	# of instances	Entropy
Cluster 1	107	0.13
Cluster 2	207	1.52
Cluster 3	202	0.33
Cluster 4	237	1.17
Cluster 5	199	0.89
Cluster 6	279	1.99
Cluster 7	279	2.20
Total Entropy		**1.34**

Table 5.4: Entropy measure for the complete C GPCR data set with the ACC transformation.

- **Digram Transformation**

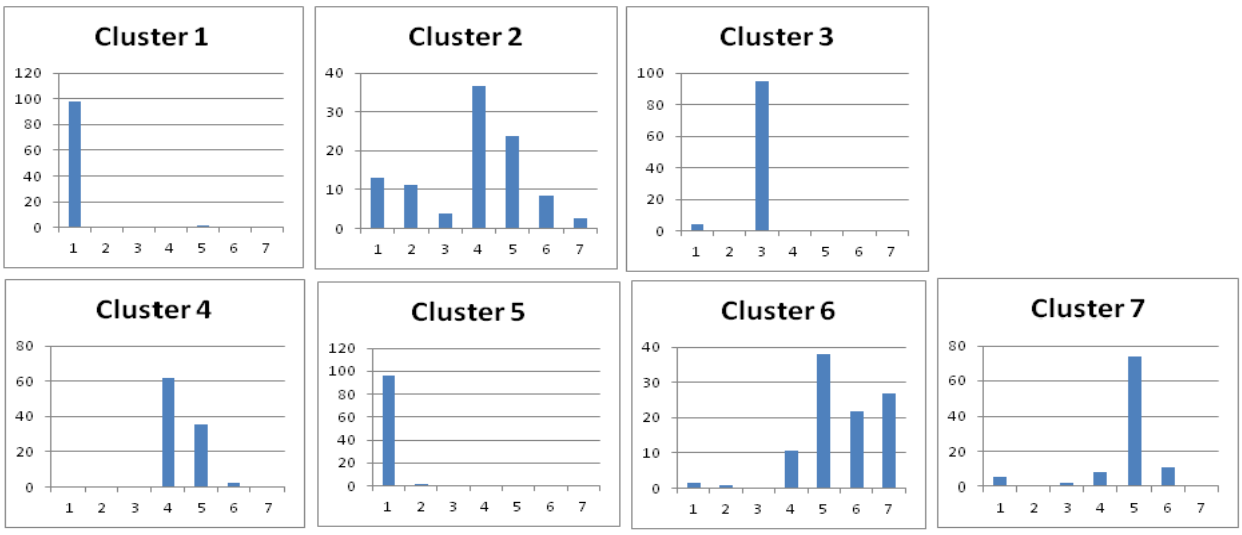

Figure 5.3: Class specificity for each cluster of the complete C GPCR data set with the Digram

transformation.

Cluster	# of instances	Entropy
Cluster 1	112	0.12
Cluster 2	374	2.39
Cluster 3	200	0.37
Cluster 4	277	1.09
Cluster 5	179	0.26
Cluster 6	205	2.02
Cluster 7	163	1.28
Total Entropy		**1.29**

Table 5.5: Entropy measure for the complete C GPCR data set with the Digram transformation.

In terms of cluster subtype-specificity, Figure 5.1 and table 5.3 show that, for the AAC data transformation, almost none of the defined clusters show clear class (subtype) specificity. Only in *cluster 1*, the first subtype (mGluR) of GPCR achieves a specificity that is close to 60%, but even in this case, the third subtype ($GABA_B$) reaches a non-negligible 30%. Several clusters show common specificity profiles: for instance, *clusters 1 and 3* are predominantly a mixture of mGluR and $GABA_B$, which means that they might truly be a single cluster with some substructure. *Cluster 4* is a very mixed combination of Pheromones and Veromonasal, but *clusters 2, 6* and *7* seem to be variations of this combination, again suggesting one main cluster with further substructure and important levels of overlapping.

The ACC and Digram transformations, instead, manage to separate some of these clusters to become more subtype-specific. mGluR and $GABA_B$ are now more clearly discriminated (*clusters 1* plus *5* and *cluster 3*, in turn) with the rest of subtypes showing clear overlapping in some clusters but also high specificity in others (for instance, Pheromones in ACC *cluster 6* and Digram in *cluster 7*).

In any case, the more complex transformations (ACC and Digram) seem to make the FCM clustering model more class C GPCR subtype-specific.

mGluR results

- **AAC Transformation**

Figure 5.4: Class specificity for each cluster of the mGluR data set with the AAC transformation.

Cluster	# of instances	Entropy
Cluster 1	32	1
Cluster 2	98	1.20
Cluster 3	126	1.29
Total Entropy		**1.22**

Table 5.6: Entropy measure for the mGluR data set with the AAC transformation.

- **ACC Transformation**

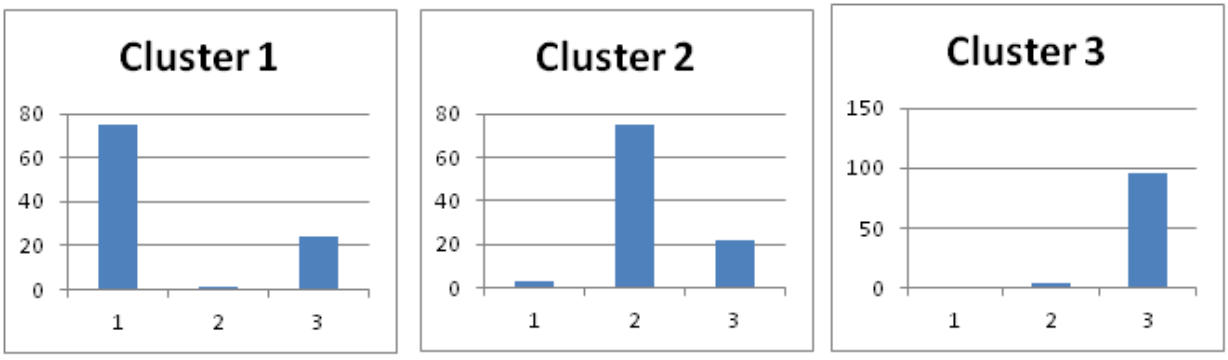

Figure 5.5: Class specificity for each cluster of the mGluR data set with the ACC transformation.

Cluster	# of instances	Entropy
Cluster 1	79	0.89
Cluster 2	72	0.94
Cluster 3	105	0.23
Total Entropy		**0.63**

Table 5.7: Entropy measure for the mGluR data set with the ACC transformation.

- **Digram Transformation**

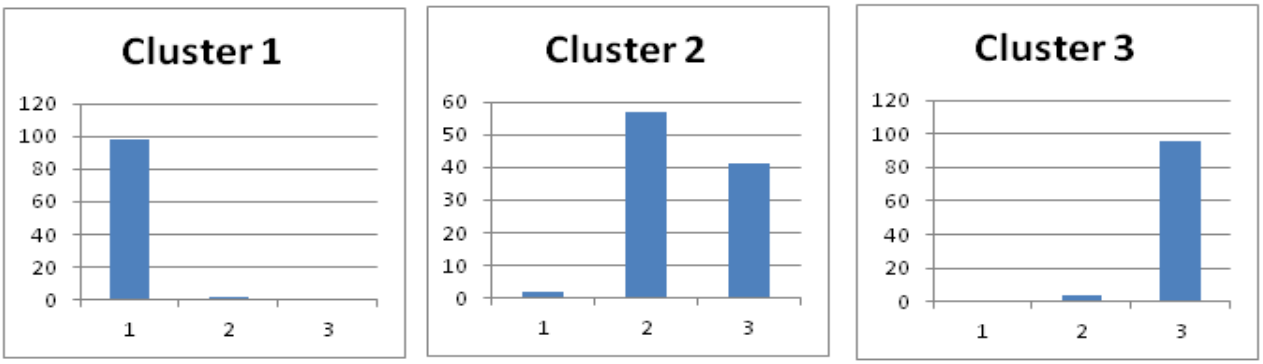

Figure 5.6: Class specificity for each cluster of the mGluR data set with the Digram transformation.

Cluster	# of instances	Entropy
Cluster 1	60	0.12
Cluster 2	95	1.11
Cluster 3	101	0.24
Total Entropy		**0.53**

Table 5.8: Entropy measure for the mGluR data set with the Digram transformation.

The corresponding results for the three groups of the mGluR subtype, reported in Figures 5.4, 5.5 and 5.6 and in Tables 5.6 5.7 and 5.8, are somehow consistent with those for the complete class C GPCR data set, which is an indication that the differences in results should mostly be attributed to the differences in the information conveyed by the different transformations. The AAC transformation cannot make the FCM cluster representation compatible with the standard grouping (*I, II* and *III*, see Table 4.2): *cluster 1* is an even mixture of Groups *II* and *III*, *cluster 2* of Groups *I* and *III* and *cluster 3* is a more mixed one somewhere in between the previous two. The three clusters have also high values of entropy as illustrated in Table 5.6.

The ACC and Digram transformations (specially the former) are more successful at naturally discriminating the groups in the FCM fully unsupervised way: each ACC cluster is very group-specific (with *cluster 3*, having a very low entropy and almost entirely corresponding to *Group III*), whereas, for Digram, *Groups I* and *III* are extremely specific of, in turn, *clusters 1* (98.3%) and *3*, but *cluster 2* is an even mixture of *Groups II* and *III*.

5.1.2 K-means for class C GPCRs

Class C GPCR results

- **AAC Transformation**

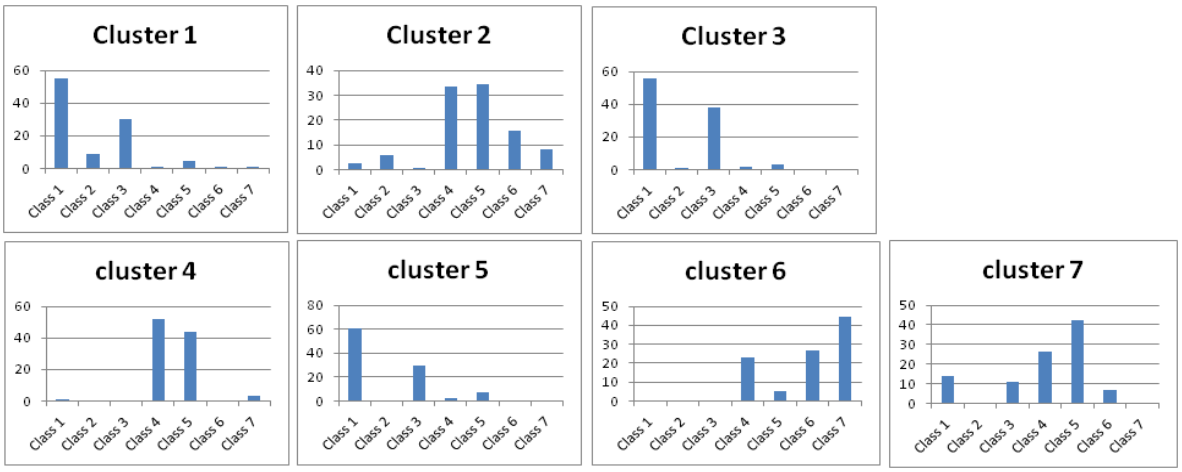

Figure 5.7: Class specificity for each cluster of the complete C GPCR data set with the AAC transformation.

Cluster	# of instances	Entropy
Cluster 1	270	1.65
Cluster 2	406	2.17
Cluster 3	196	1.33
Cluster 4	189	1.24
Cluster 5	54	1.34
Cluster 6	56	1.74
Cluster 7	339	2.03
Total Entropy		**1.77**

Table 5.9: Entropy measure for the complete C GPCR data set with the AAC transformation.

- **ACC Transformation**

Figure 5.8: Class specificity for each cluster of the complete C GPCR data set with the ACC transformation.

Cluster	# of instances	Entropy
Cluster 1	260	0.26
Cluster 2	136	0.72
Cluster 3	150	0.23
Cluster 4	188	1.07
Cluster 5	165	1.57
Cluster 6	379	1.79
Cluster 7	232	1.97
Total Entropy		**1,19**

Table 5.10: Entropy measure for the complete C GPCR data set with the ACC transformation.

- **Digram Transformation**

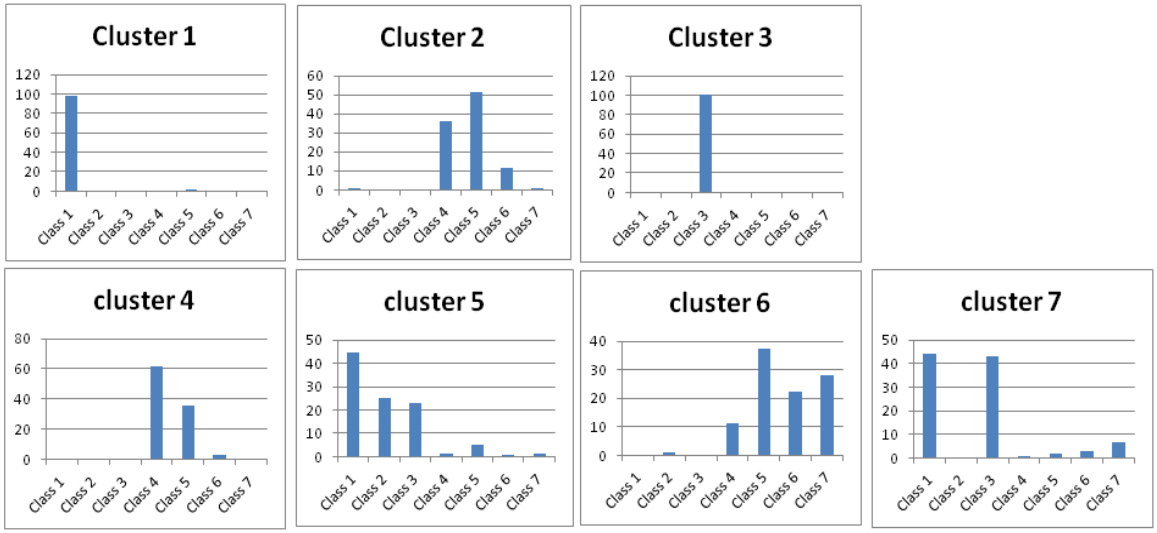

Figure 5.9: Class specificity for each cluster of the complete C GPCR data set with the Digram transformation.

Cluster	# of instances	Entropy
Cluster 1	222	0.10
Cluster 2	398	1.47
Cluster 3	121	0
Cluster 4	284	1.10
Cluster 5	184	1.90
Cluster 6	197	1.95
Cluster 7	104	1.63
Total Entropy		**1.39**

Table 5.11: Entropy measure for the complete C GPCR data set with the Digram transformation.

The results of K-means algorithm for the seven subtypes of class C GPCRs, reported in Figures 5.7, 5.8 and 5.9 as well as in Tables 5.9, 5.10 and 5.11, are consistent with those of FCM algorithm. Again with AAC data transformation, almost none of the defined clusters show clear class (subtype) specificity. We can also observe that *clusters 1 and 3* are a mixture of mGluR and $GABA_B$, as in FCM results, which means that they might truly be a single cluster with some substructure. Also, cluster 4 is a combination of Pheromones and Veromonasal, and *clusters 2, 6* and *7* are variations of this combination, as in FCM, again suggesting one main cluster with further substructure and important levels of overlapping.

The ACC and Digram transformation manage to separate some of these clusters to become more subtype-specific. The similarity of the two algorithms is that mGluR and $GABA_B$ are now more clearly discriminated from the rest subtypes. Moreover, in the ACC transformation, Pheromone can also be discriminated from the rest subtypes due to the high specificity that has in cluster 2. The rest of the subtypes show clear overlapping in some of the cluster.

Comparing the results of both algorithms in terms of the total entropy measure, the conclusions are not clear-cut. ACC and Digram show a clear advantage both in FCM and K-Means, but neither algorithm shows a clear advantage over the other.

mGluR results

- **AAC Transformation**

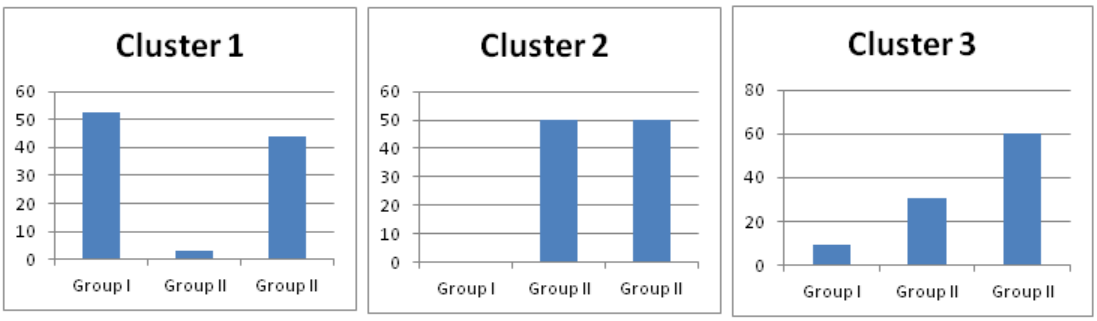

Figure 5.10: Class specificity for each cluster of the mGluR data set with the AAC transformation.

Cluster	# of instances	Entropy
Cluster 1	93	1.17
Cluster 2	32	1
Cluster 3	131	1.28
Total Entropy		**1.20**

Table 5.12: Entropy measure for the mGluR data set with the AAC transformation.

- **ACC Transformation**

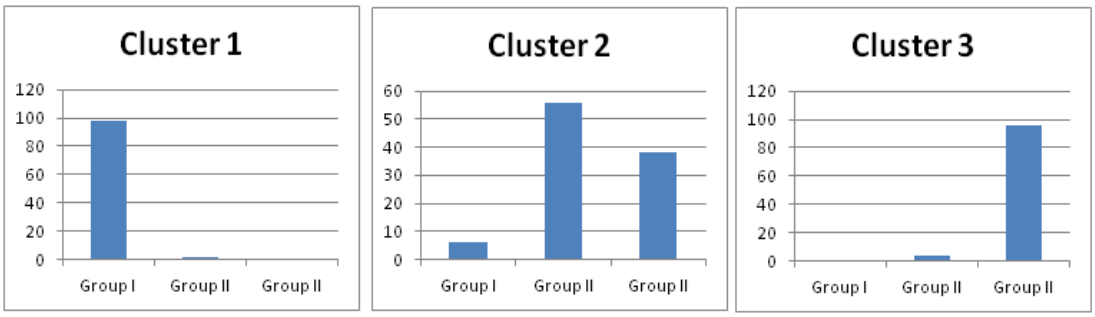

Figure 5.11: Class specificity for each cluster of the mGluR data set with the ACC transformation.

Cluster	# of instances	Entropy
Cluster 1	56	0.13
Cluster 2	97	1.25
Cluster 3	103	0.24
Total Entropy		**0.59**

Table 5.13: Entropy measure for the mGluR data set with the ACC transformation.

- **Digram Transformation**

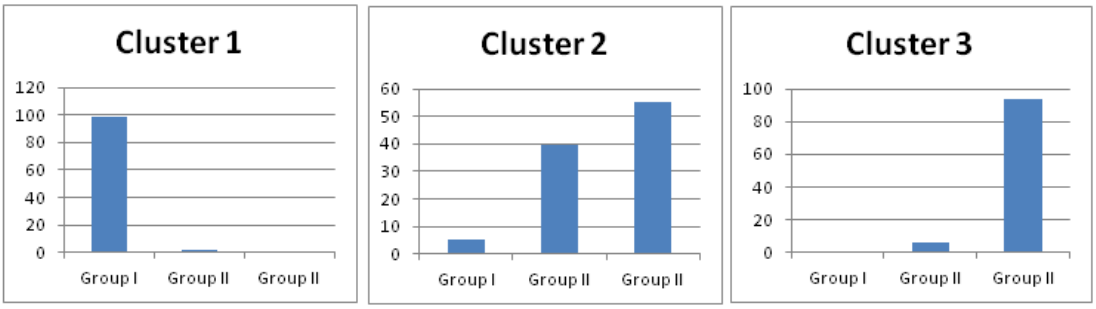

Figure 5.12: Class specificity for each cluster of the mGluR data set with the Digram transformation.

Cluster	# of instances	Entropy
Cluster 1	55	0.13
Cluster 2	136	1.22
Cluster 3	65	0.33
Total Entropy		**0.76**

Table 5.14: Entropy measure for the mGluR data set with the Digram transformation.

The corresponding results of K-means for the three groups of the mGluR subtype, illustrated in figures 5.10, 5.11 and 5.12 and in Tables 5.12, 5.13 and 5.14 are, overall, equally similar to the FCM results. The AAC transformation cannot discriminate any of the standard mGluR groupings, with all clusters yielding high entropy values. On the other hand, for both the AAC and Digram transformations, *Groups I* and *III* are extremely specific of, in turn, *cluster 1* and *cluster 3* but in both transformations *cluster 2* is a mixture of the three groups.

Comparing the results of the algorithms in terms of the total entropy measure, the conclusions are again mixed. ACC and Digram show a clear advantage both in FCM and K-Means (consistent with the class-specificity) and, again, Digram provides the best result for FCM, while ACC provides the best one for K-Means.

The results obtained from FCM and K-means algorithms give us a clear indication that the differences in results should mostly be attributed to the differences in the information conveyed by the different sequence transformations. AAC cannot neatly discriminate the different subtypes of class C GPCRs, whereas the obtained clusters are more subtype specific for the richer ACC and Digram transformations.

5.2 Experiments with varying number of clusters: Cluster Stability Analysis

In the previous section, the six datasets illustrated in Table 5.1 were applied to FCM and K-means algorithms, using a fixed number of clusters that corresponded to the known receptor subtypes provided by GPCRDB. In this section, this constraint is removed and the experiments are focused on the analysis of the stability of the results of each clustering algorithm.

As already explained, recent experimental evidence [65] has shown that K-Means solutions that might be expected to be similar according to the final value of the objective function, may in fact be quite dissimilar, suggesting the convenience of using the objective function as a criterion of model optimality only in combination with a cluster stability criterion for cluster partition reproducibility such as the SeCo maps introduced in the previous chapter. We have proposed an extension of these maps to FCM by first defining

weighted contingency tables and a corresponding weighted Cramér's V index. In the following experiments, we aim to assess the usability of SeCo maps with FCM and the extent to which FCM suffers the same problems as K-Means in terms of increasing instability as the value of *K* increases.

For the experimental results presented in the following sections, nin was set to a value of 500. Moreover, for class C GCPR transformed data sets, the range of the values of K (number of clusters) was varied from 2 to 12, while for the mGluR transformed data sets, it was varied from 2 to 10. For each one of the data sets, three SeCo maps were created using the following:

1. The K-means objective function and the standard Cramér's V index calculation
2. The FCM objective function and the standard Cramér's V index calculation
3. The FCM objective function and the novel weighted Cramér's V index calculation proposed in the previous chapter.

In section 5.2.1, the three SeCo maps for each dataset and each algorithm are presented, while, in section 5.2.2, and according to recommendations in [65], a threshold was applied to the ΔSSQ values so that only the 10% top values of ΔSSQ were used to create the SeCo maps (this 10% choice is expected to allow the degeneracy of similar SSQ values to be resolved by choosing the individual cluster partition with maximum value of the internal consistency index).

5.2.1 Full Separation Concordance (SeCo) map

5.2.1.1 Class C GPCR results

AAC Transformation

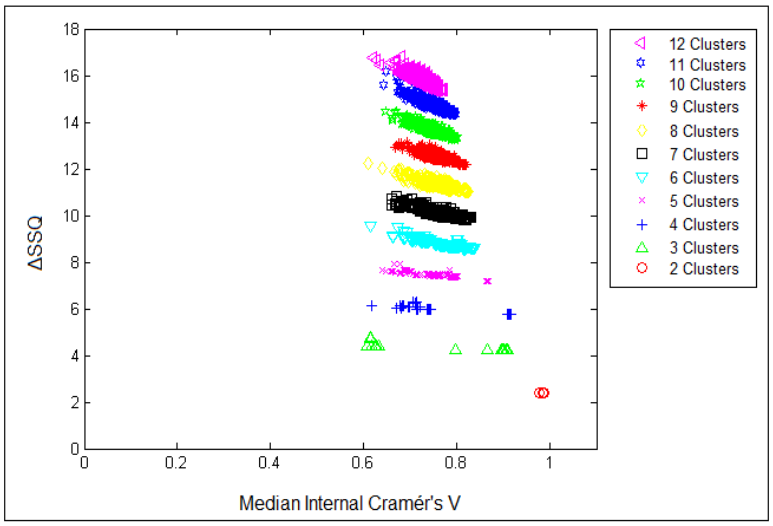

Figure 5.13: Separation Corcondance map for CGPCR_AAC dataset. ΔSSQ on the y-axis and the median Cramérs' V on the x-axis, for 500 initializations for each value of c, from 2 to 12, for K-Means.

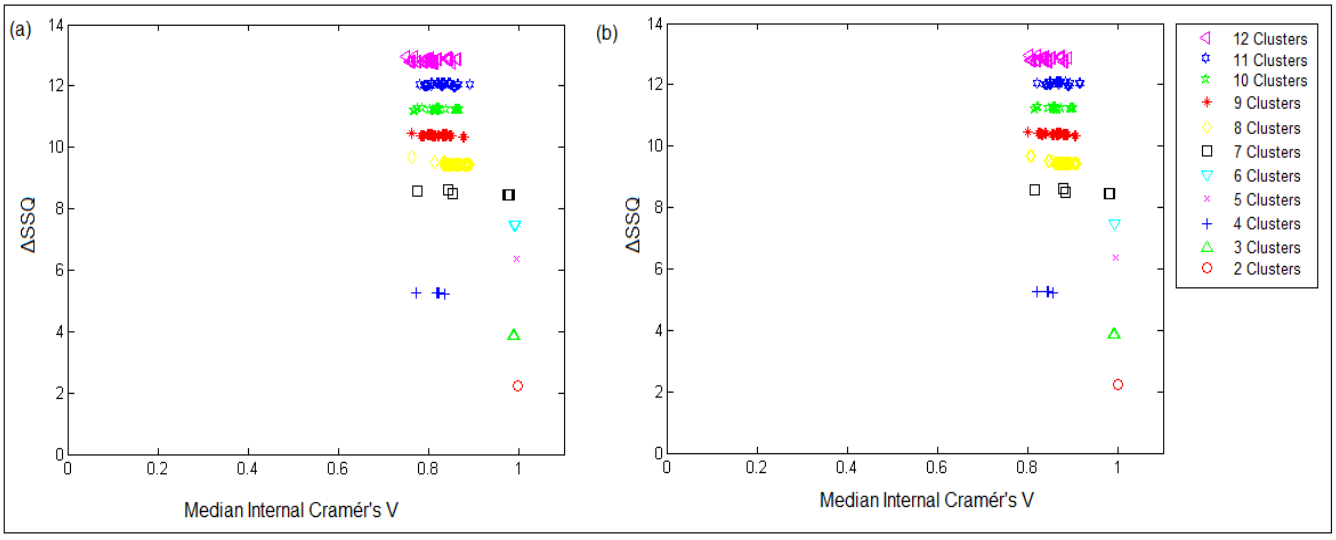

Figure 5.14: Separation Corcondance map for CGPCR_AAC dataset. Representation as in previous figure; (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index..

ACC Transformation

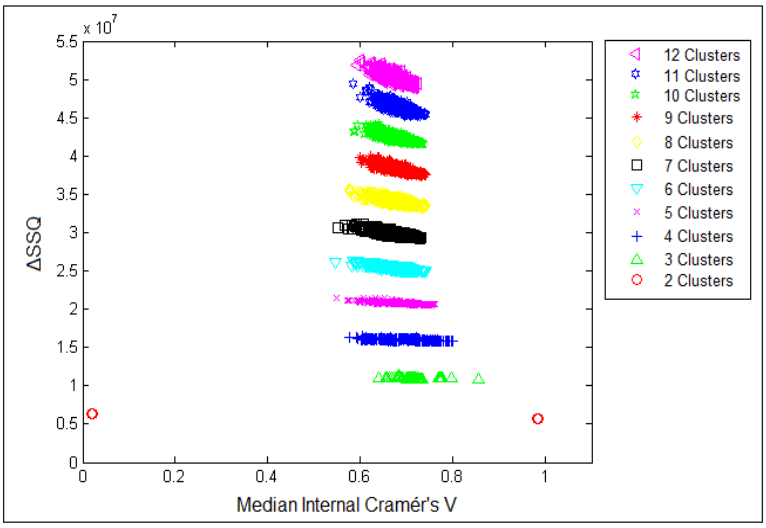

Figure 5.15: Separation Corcondance map for CGPCR_ACC dataset. ΔSSQ on the y-axis and the median Cramérs' V on the x-axis, for 500 initializations for each value of c, from 2 to 12, for K-Means.

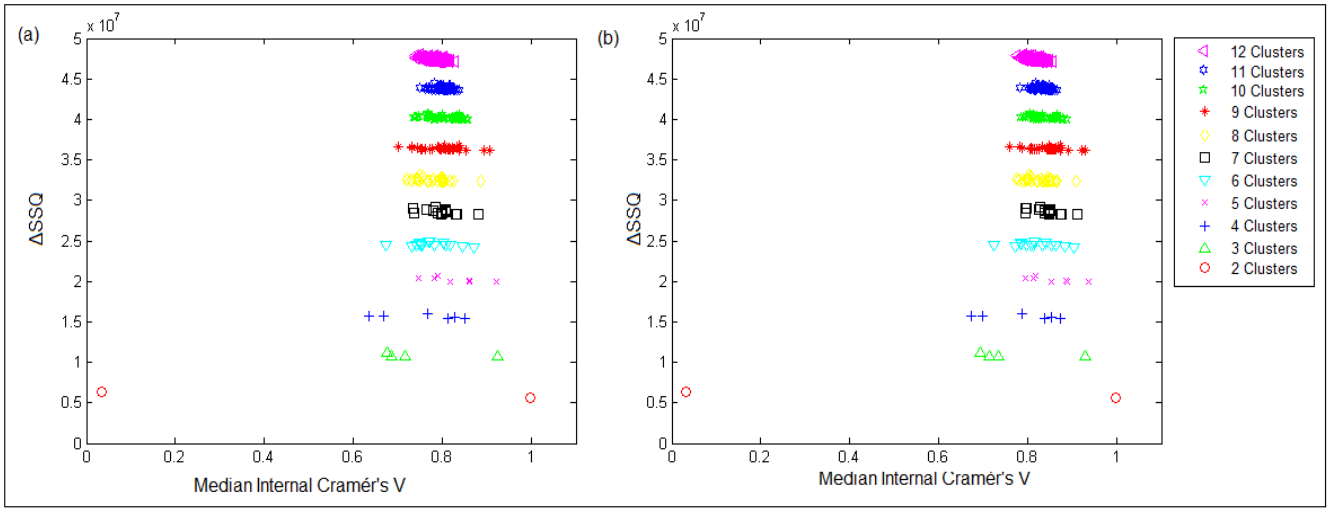

Figure 5.16: Separation Corcondance map for CGPCR_ACC dataset. Representation as in previous figure; (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index..

Digram Transformation

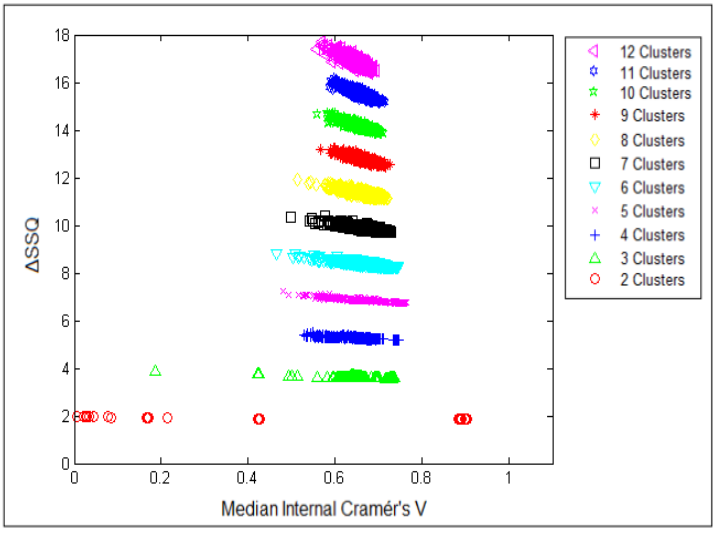

Figure 5.17: Separation Corcondance map for CGPCR_Digram dataset. ΔSSQ on the y-axis and the median Cramérs' V on the x-axis, for 500 initializations for each value of c, from 2 to 12, for K-Means.

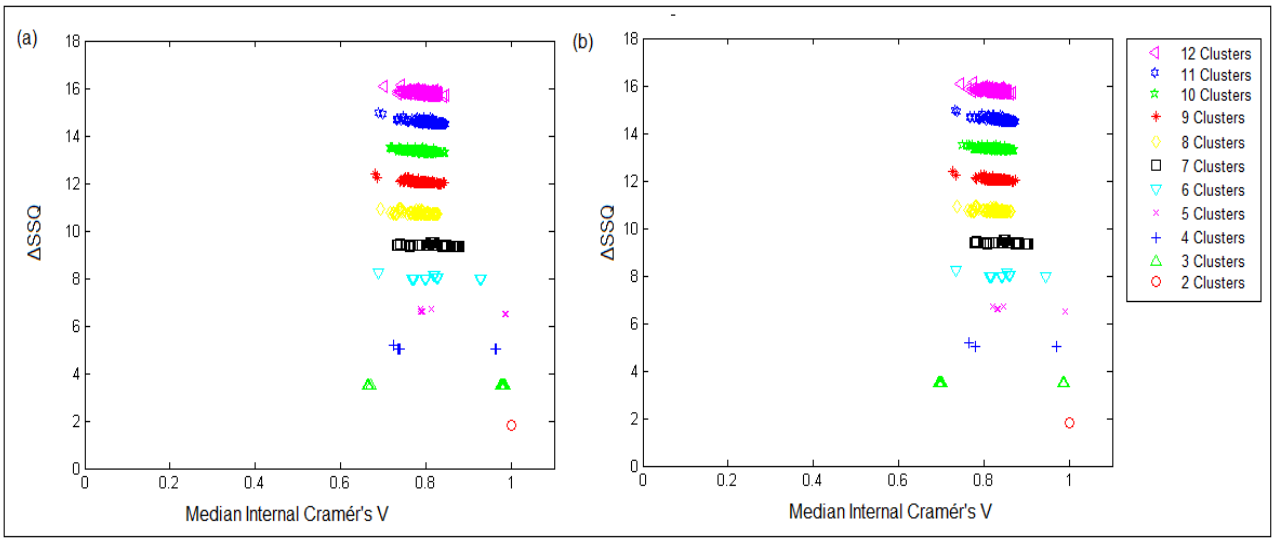

Figure 5.18: Separation Corcondance map for CGPCR_Digram dataset. Representation as in previous figure; (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index.

The complete SeCo maps in Figures 5.13 to 5.18 for the full class C GPCR data set reveal some interesting insights on clustering algorithm stability and its relation to the ΔSSQ values for different values of parameter K.

First, and partially corroborating the results reported in [65], the K-Means algorithm is shown to suffer from a wide spread on stability (as measured by the median Cramér's V index) for solutions with a very similar value of ΔSSQ. Interestingly, this effect does not necessarily increase as K increases, as also reported in [65], for any of the data transformations. In fact, the variability somehow decreases. This is likely to be the result of highly overlapping clusters, with plenty of substructure.

Second, the FCM algorithm yields far more stable results than K-Means, with a much more limited spread (median Cramér's V index results clustered in a more reduced number of values). These results are also very consistent over data transformations. Overall, this indicates that FCM is much more resilient than its crisp K-Means counterpart to the variability introduced by random initializations.

Third, the stability results as measured by the standard Cramér's V index and the proposed *weighted* Cramér's V index are strikingly similar for all data transformations, providing support for the use of the latter, which is a more faithful account of the true belief of the algorithm regarding cluster membership.

Fourth and final, the SeCo map was proposed in [65] as a method that could provide as with relative guidance with the regard to the most adequate value of K supported by the data. This is very unclear from the reported results, as no value of K provides a differentially better combination of ΔSSQ and stability. This is particularly true for K-Means, but not too different for FCM. This would suggest, as in [65], that the segregated analysis of the best 10% results might be an adequate alternative.

5.2.1.2 mGluR results

AAC Transformation

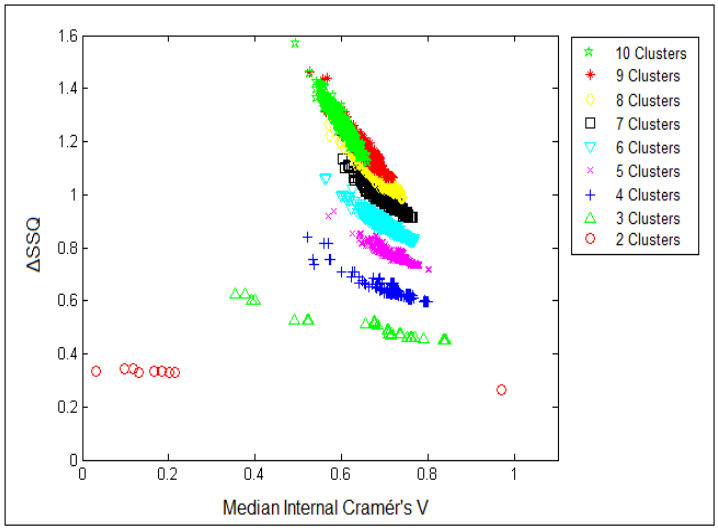

Figure 5.19: Separation Corcondance map for mGluR_AAC dataset. ΔSSQ on the y-axis and the median Cramérs' V on the x-axis, for 500 initializations for each value of c, from 2 to 10, for K-Means.

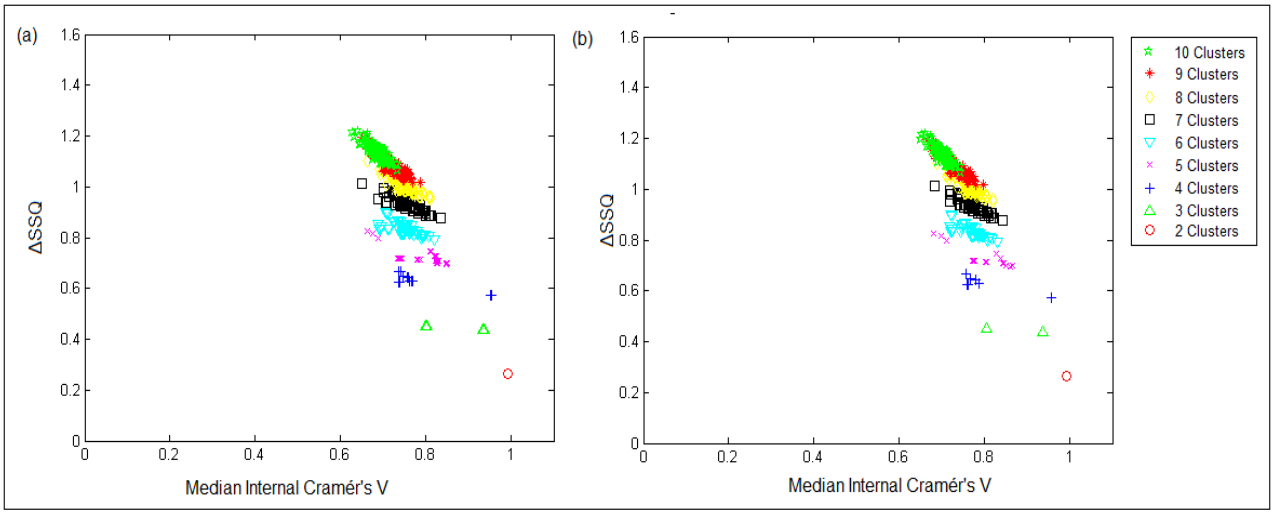

Figure 5.20: Separation Corcondance map for mGluR_AAC dataset. Representation as in previous figure; (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index.

ACC Transformation

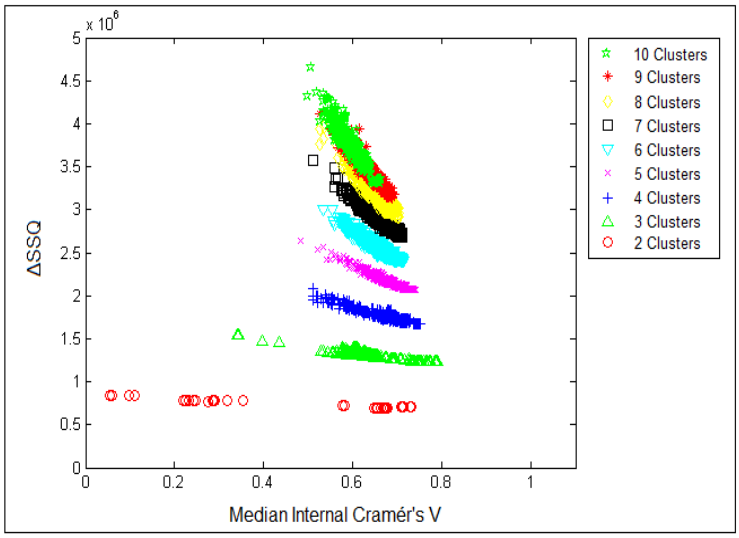

Figure 5.21: Separation Corcondance map for mGluR_ACC dataset. ΔSSQ on the y-axis and the median Cramérs' V on the x-axis, for 500 initializations for each value of c, from 2 to 10, for K-Means

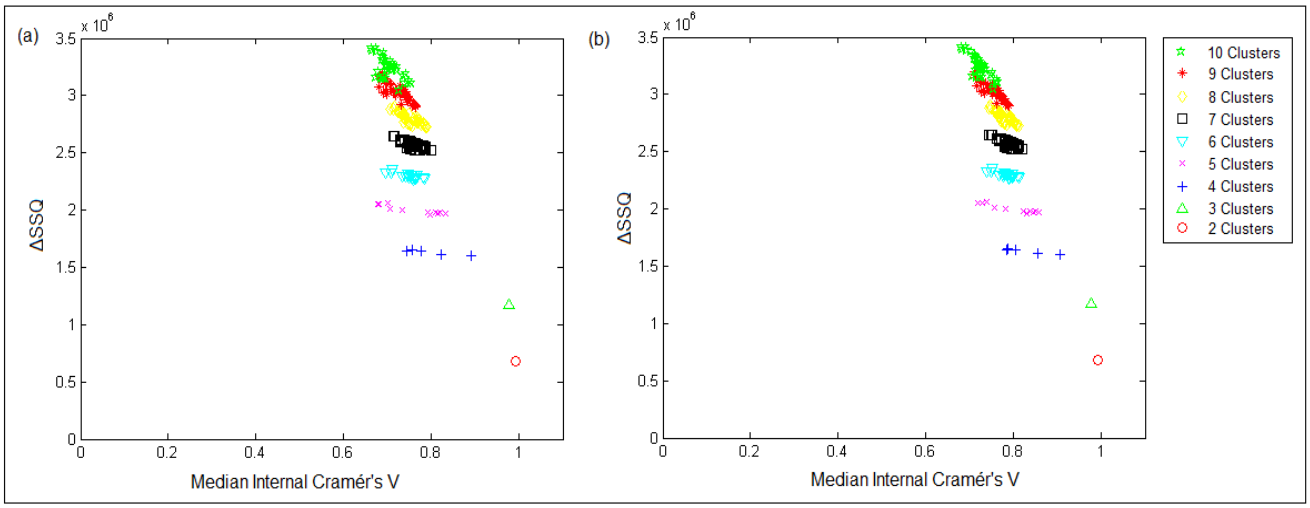

Figure 5.22: Separation Corcondance map for mGluR_ACC dataset. Representation as in previous figure; (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index.

Digram Transformation

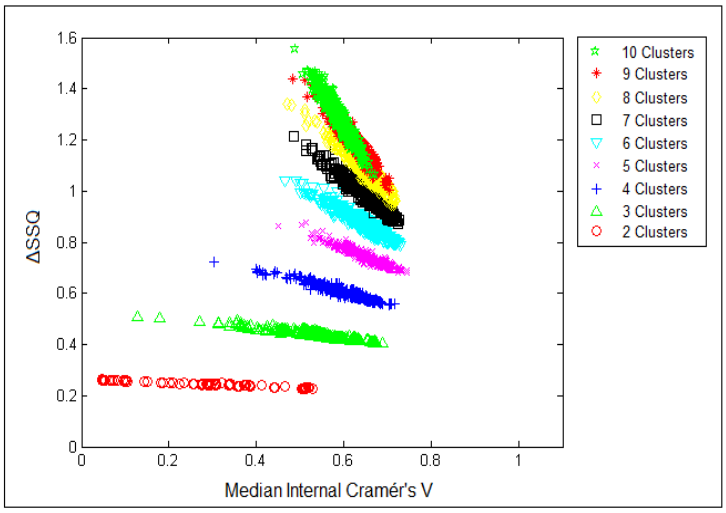

Figure 5.23: Separation Corcondance map for mGluR_ Digram dataset. ΔSSQ on the y-axis and the median Cramérs' V on the x-axis, for 500 initializations for each value of c, from 2 to 10, for K-Means

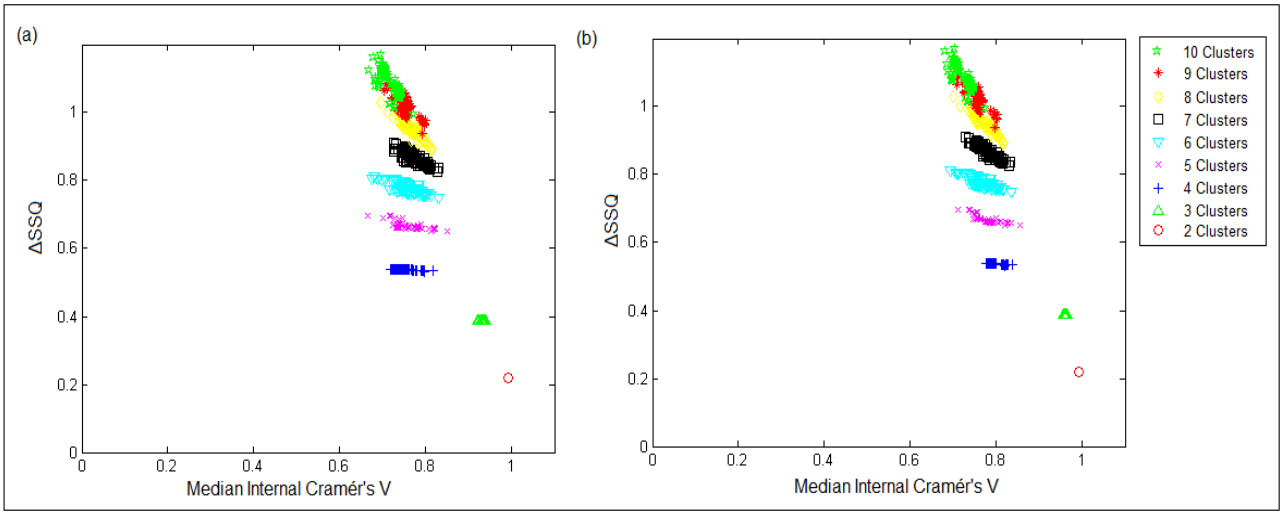

Figure 5.24: Separation Corcondance map for mGluR_Digram dataset. Representation as in previous figure; (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index.

The complete SeCo maps in Figures 5.19 to 5.24 for the mGluR data subset is overall consistent with the previously discussed results for the complete class C GPCR data set.

First, and again partially corroborating the results reported in [65], the K-Means algorithm is shown to suffer from a wide spread on stability for solutions with the same K and that this effect clearly decreases this time as K increases for any of the data transformations. It is also true, though, that, as K increases, the value of ΔSSQ remains increasingly dissimilar for solutions of the same K.

Second, again the FCM algorithm yields far more stable results than K-Means, with a much more limited spread over data transformations. Although this confirms that FCM is more resilient than K-Means to the variability introduced by random initializations, we now see that low values of K yield almost identical stability *vs.* ΔSSQ. See, for instance, that the standard subtyping of mGluRs was established in three types (*I, II* and *III*), while the SeCo maps suggest that this is a differentially stable solution.

Third, little difference is to be found between the standard Cramér's V index and the proposed *weighted* Cramér's V index.

5.2.2 Thresholding the objective function

A threshold was subsequently applied to the ΔSSQ values, such that only the top 10% of the values are now considered for the creation of the SeCo map. According to [65], the use of a threshold exposes the true underlying partitions of the data. A threshold of 10% was chosen to be used, because as it was shown by the experiments in [65], as the threshold is tightened the consistency of SeCo maps was improving.

The obtained results are presented below, and again we have three SeCo maps for each dataset as described in the previous section.

5.2.2.1 Class C GPCR results

AAC Transformation

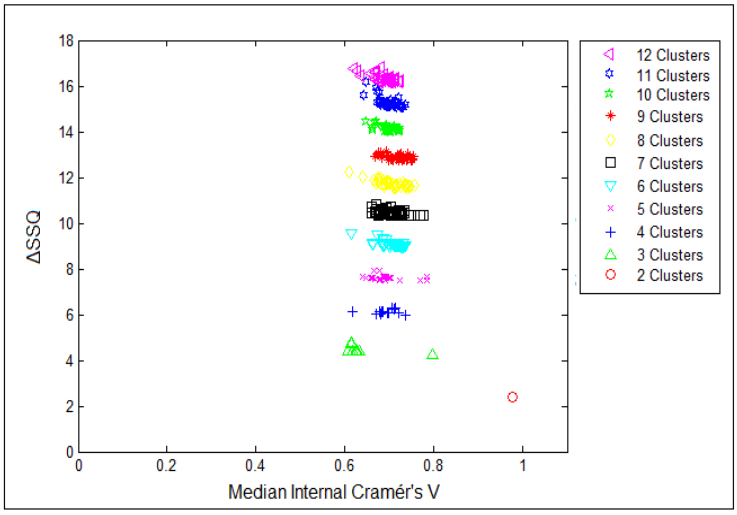

Figure 5.25: Separation Corcondance map for CGPCR_AAC dataset using 10% threshold for separation metric. K-Means. Representation as in previous figures.

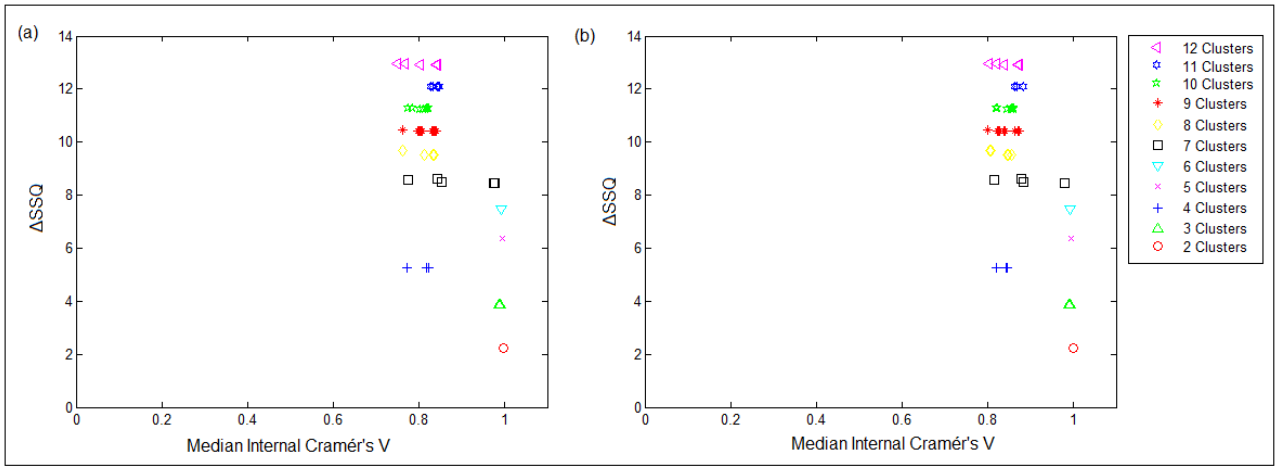

Figure 5.26: Separation Corcondance map for CGPCR_AAC dataset using 10% threshold for separation metric. (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index. Representation as in previous figures.

ACC Transformation

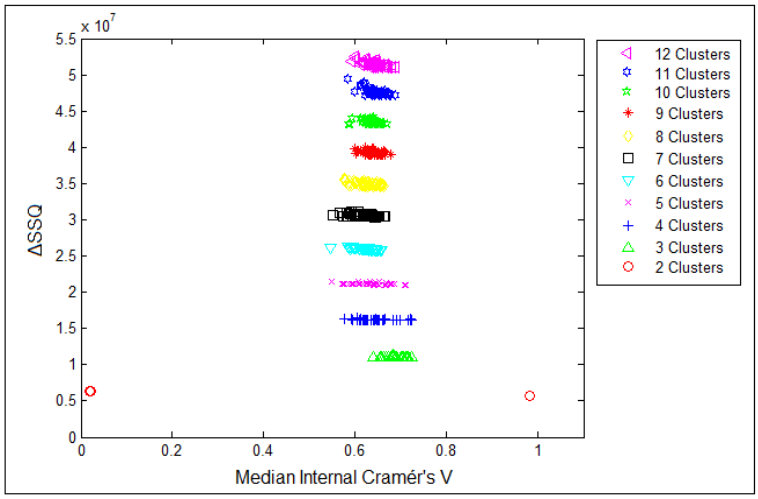

Figure 5.27: Separation Corcondance map for CGPCR_ACC dataset using 10% threshold for separation metric. K-Means. Representation as in previous figures.

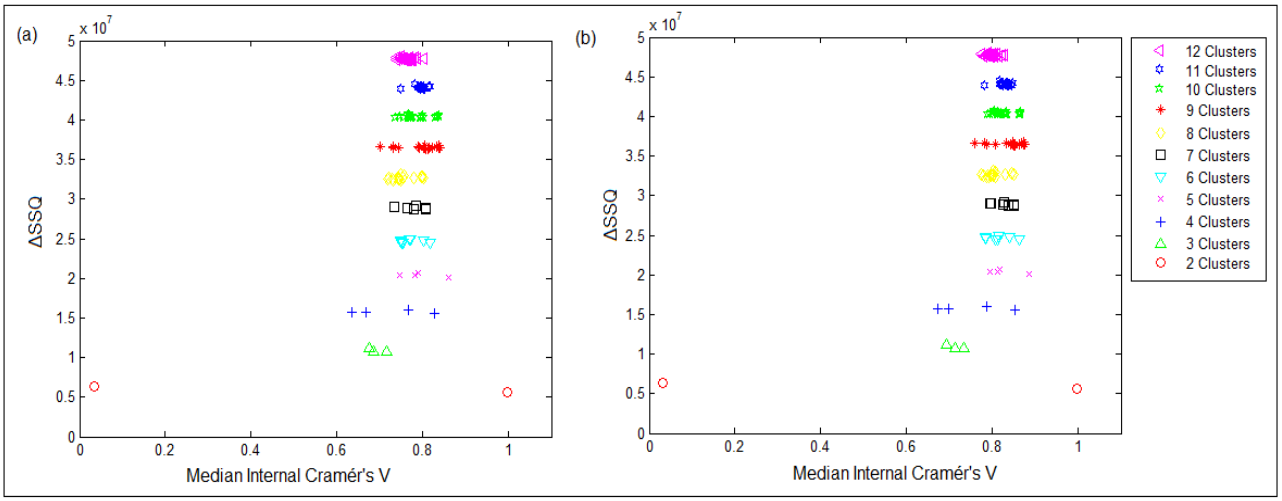

Figure 5.28: Separation Corcondance map for CGPCR_ACC dataset using 10% threshold for separation metric. (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index. Representation as in previous figures.

Digram Transformation

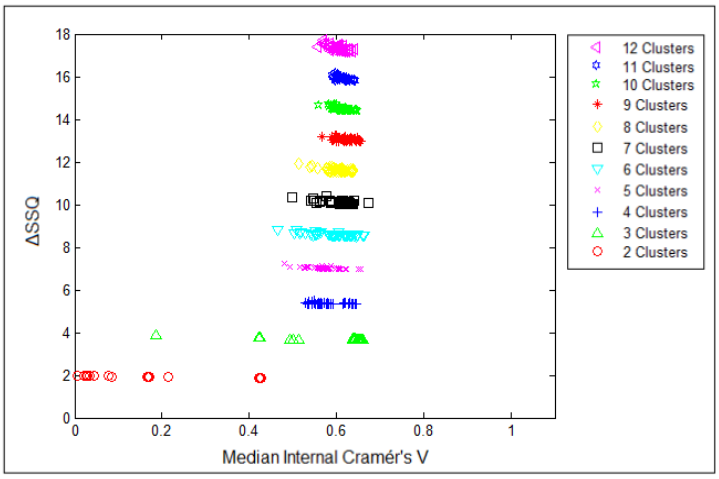

Figure 5.29: Separation Corcondance map for CGPCR_ Digram dataset using 10% threshold for separation metric. K-Means. Representation as in previous figures.

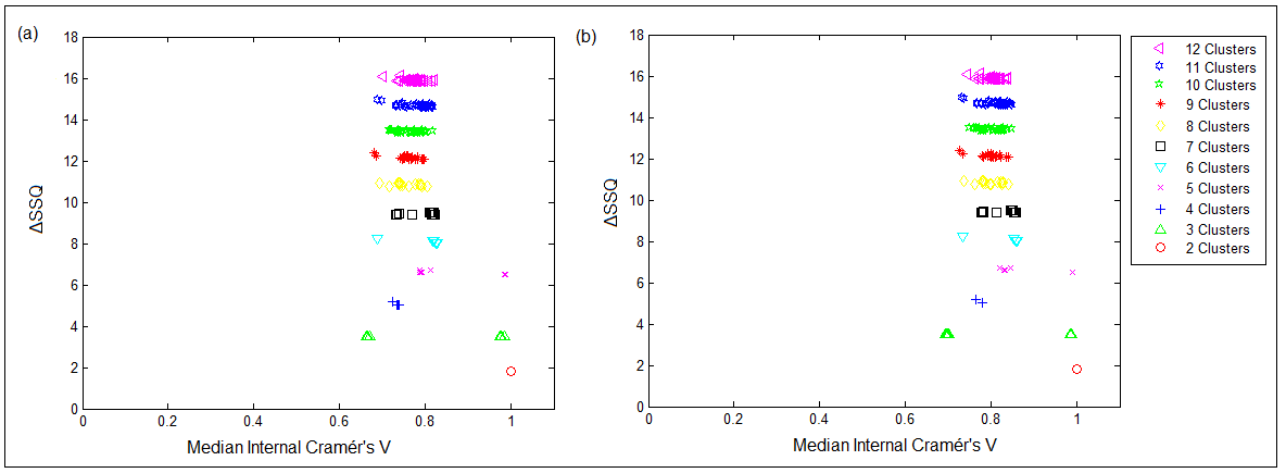

Figure 5.30: Separation Corcondance map for CGPCR_Digram dataset using 10% threshold for separation metric. (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index. Representation as in previous figures.

5.2.2.2 mGluR results

AAC Transformation

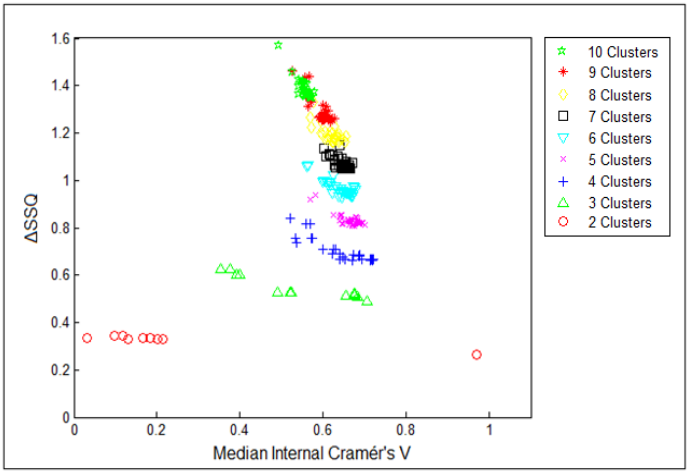

Figure 5.31: Separation Corcondance map for mGluR_AAC dataset using 10% threshold for separation metric. K-Means. Representation as in previous figures.

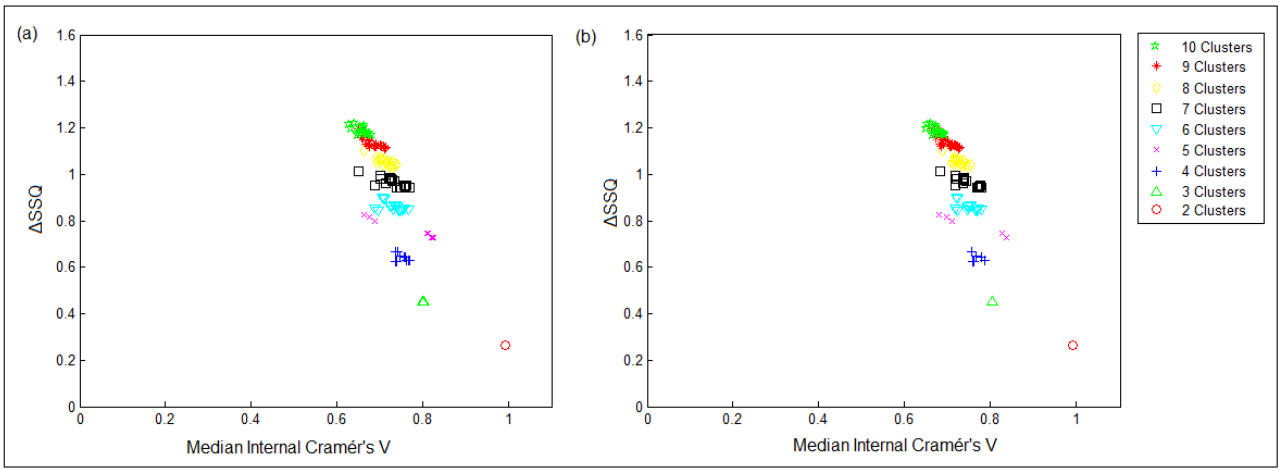

Figure 5.32: Separation Corcondance map for mGluR_AAC dataset using 10% threshold for separation metric. (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index. Representation as in previous figures.

ACC Transformation

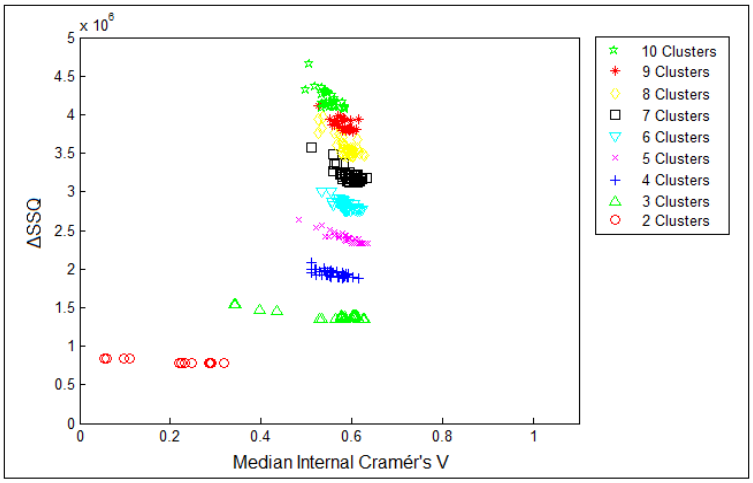

Figure 5.33: Separation Corcondance map for mGluR_ACC dataset using 10% threshold for separation metric. K-Means. Representation as in previous figures.

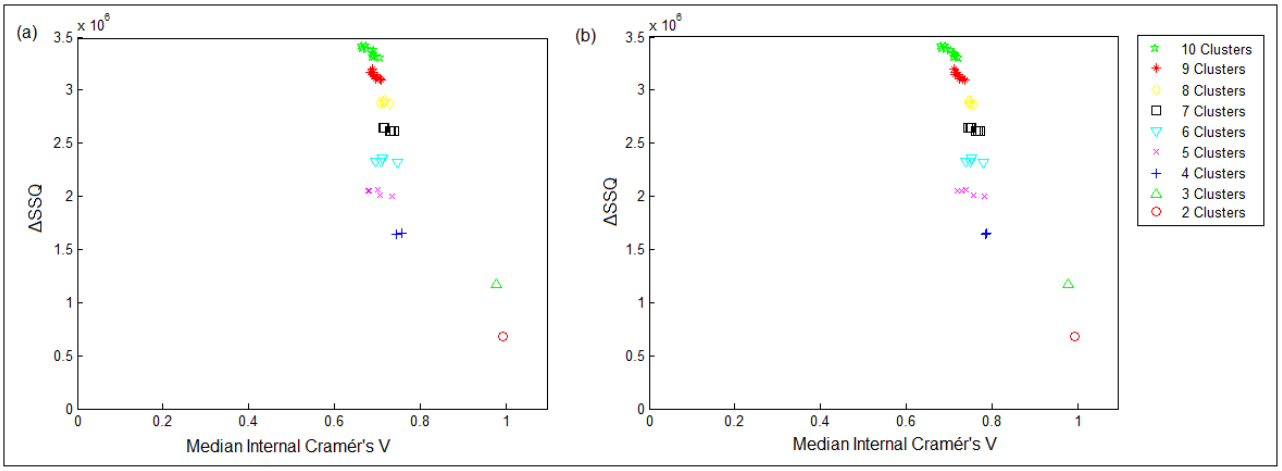

Figure 5.34: Separation Corcondance map for mGluR_ACC dataset using 10% threshold for separation metric. (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index. Representation as in previous figures.

Digram Transformation

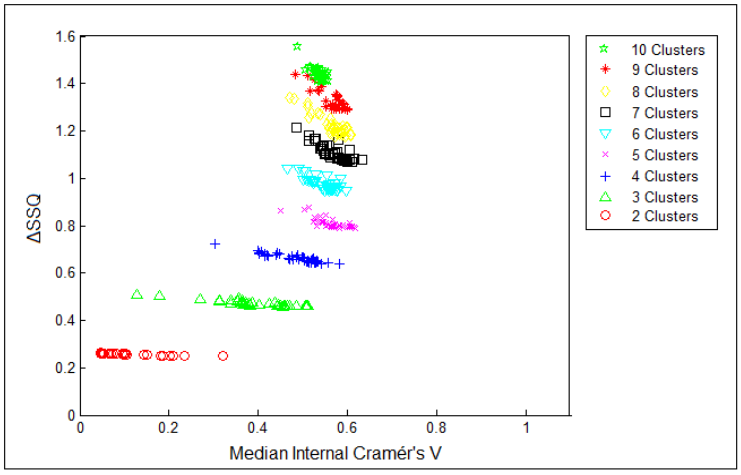

Figure 5.35: Separation Corcondance map for mGluR_ Digram dataset using 10% threshold for separation metric. K-Means. Representation as in previous figures.

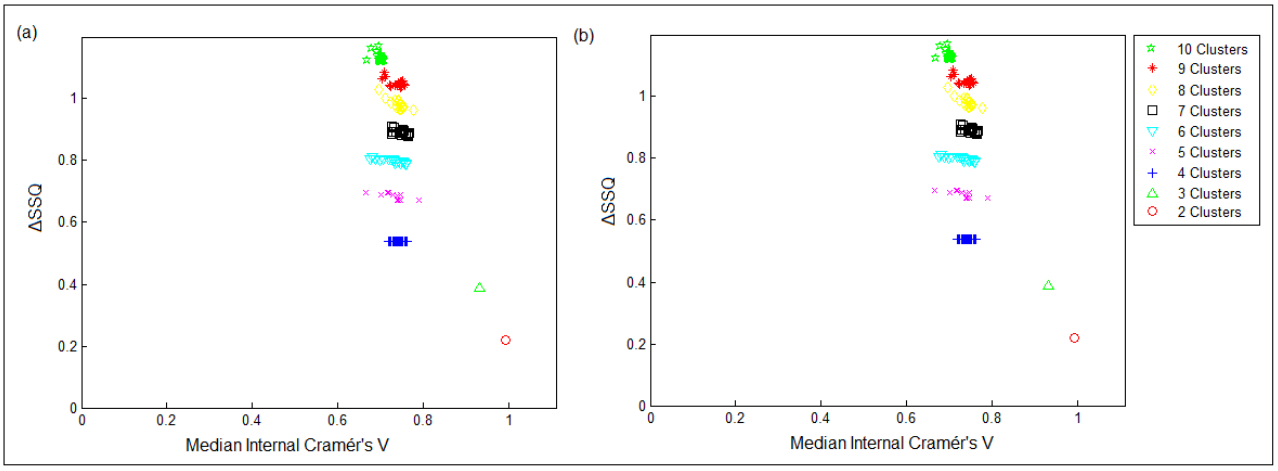

Figure 5.36: Separation Corcondance map for mGluR_ Digram dataset using 10% threshold for separation metric. (a) FCM with standard Cramér's V index; (b) FCM with weighted Cramér's V index. Representation as in previous figures.

As previously mentioned, a threshold for the ΔSSQ values to select the 10% top values for each value of K is expected to allow the degeneracy of similar SSQ values to be resolved by choosing the individual cluster partition with maximum value of the internal consistency index.

First, these FCM results, as reported in Figs. 5.25 to 5.34, are clearly much more parsimonious than the complete ones, revealing a high concentration of stability results around just a handful of values, in comparison with the still wide spread of K-Means. For K-Means, this effect does not necessarily increase as K increases for any of the data transformations. For FCM, though, an increase in spread as K increases is revealed.

Second, the stability results as measured by the standard Cramér's V index and the proposed *weighted* Cramér's V index are again very similar for all data transformations (even better for the latter, providing further support for the proposed method).

Third, the restricted 10% SeCo maps are now a slightly better guide to make a decision about the most adequate value of K, as supported by the data. For the complete class C GPCR data set as analyzed by FCM, a solution beyond 7 clusters is clearly not supported by the AAC transformation, as maximum stability suddenly decreases at the same point the cluster model becomes more unstable (with more spread values). Note that this corresponds with the "natural" description of subtypes for this data set. A similar conclusion is not supported for the ACC transformation and only partially for Digram, whose low-K solutions are clearly polarized. For the mGluR data set, up to a 5-cluster solution is supported by the AAC transformation, whereas 3 or 4-cluster solutions seem to be preferred for ACC and Digram. In any case, these results are hardly conclusive, which means that SeCo maps have limited applicability for the choice of K in highly overlapping data sets such as those analyzed in this thesis.

Chapter 6

Conclusions and Future Work

6.1 Conclusions 84

6.2 Future Work 86

6.1 Conclusions

GPCRs are cell membrane proteins with a very relevant role in many biological processes. They have become a target of intensive research due to their impact in the field of pharmacology. In particular, class C of GPCRs regulates a number of important physiological functions and thus they are intensively pursued as drug targets.

The 3-D structure of most GPCRs is unknown, and this is especially true for class C of this superfamily, the goal of this thesis. The investigation of their functionality is therefore often limited to the analysis of their primary amino acid sequences. The unraveling of the three-dimensional structure of GPCRs, will determine its ability for certain ligand binding, and therefore, their applicability in pharmacological research. The discrimination of GPCRs based on their primary amino acid sequences is a building block towards such full characterization.

In the reported research, a fuzzy clustering technique, namely FCM, was selected to be the main tool for the grouping structure and subtype discriminability analyses of class C GPCRs and, more in particular, also mGlu receptors, from different transformations of their unaligned amino acid sequences, based both on residue frequencies and on their physicochemical properties. The standard K-Means algorithm was also used in some of the experiments as a well-known benchmark method against which to compare the performance of FCM.

We have also paid particular attention to an analysis of the stability of the clustering algorithms. Assuming that clustering solutions that strike a balance between low error and high stability ought to be sought, and following reports that claim that similar K-means solutions in terms of the error may in fact be quite dissimilar, we have explored the convenience of using the objective function as a criterion of model optimality in combination with stability criteria for cluster partition reproducibility. Particularly, we have used Separation / Concordance maps, originally defined for crisp clustering, and which we have generalized to cover also fuzzy and probabilistic clustering techniques.

The experimental results, described in the previous chapter, have shown that K-means and FCM clustering models are able to provide an unsupervised partition of the complete C GPCR analyzed data set into clusters that show different levels of subtype specificity. Some subtypes such as mGluR (of special interest in pharma) and $GABA_B$ are easier to discriminate in this way than the rest of subtypes, providing us with an idea of the extent of the discriminability of these subtypes on the basis of different data transformations. The obtained clusters form both algorithms are more subtype specific for the more complex ACC and Digram transformations than for the AAC transformation.

For the mGluR datasets all the transformations yield much more clear-cut, which implies that even reasonably simple alignment-free transformations of the primary protein sequences are consistent with the standard partition of the mGluR family into three main groups that, in turn, reorganize their eight recognized subtypes.

Moreover, the experimental results for cluster stability have shown that FCM has overall higher stability than K-means for results of similar error using the same number of clusters (K). The FCM results for the standard and weighted Cramérs' V index are much more compact and consistent over the value of K than K-means results, which have a wider spread of the median Cramérs' V index in their corresponding SeCo maps.

Also (and interestingly) the FCM results obtained using the proposed weighted Cramérs' V index calculation yield similar or better stabilities than the ones using the standard Cramérs' V index. This indicates that for the FCM algorithm, it is more suitable to use the weighted Cramérs' V index for the SeCo map creation, because it fully reflects the

output of the FCM algorithm.

Furthermore, the SeCo maps reported in the previous chapter have not shown an increase of the spread of the stability as the value of K was increased, as was expected. We consider this to be further evidence that the data sets analyzed have an intrinsically high level of overlapping and subtype substructure. Thus, due to the nature of the data sets, the SeCo map results do not provide a strong indication for the selection of the adequate number of clusters for each data set.

6.2 Future Work

A straightforward extension of the FCM clustering model for future research is precisely its definition within a hierarchical framework. The investigation of class C GPCR subtypes at deeper level of sub-typing can provide information about the homology of each subtype as well as about the different levels of sub-subtypes that may exist. The existence of true substructure on the analyzed data is documented in proteomics literature and is clearly hinted by the nature of the results reported in this thesis.

In future research, FCM clustering might be used for the investigation of receptors with very heterogeneous grouping structure. This heterogeneity might be a clue to their susceptibility towards heterodimerization, which could be useful in the quest of more potent and safer drugs. The FCM could also offer the possibility of detecting receptors with mixed membership values (indicating that the certainty that this receptor belongs to a specific cluster is low) for further analysis.

Research using fuzzy models could also be specialized to analyze not complete primary sequences, but fragments of these sequences such as the intracellular and extracellular domains, as well as the 7TM domain. They could also focus on *motifs* (sequence fragments of known interest) previously reported in the literature of the field.

Furthermore, the method proposed in this thesis can be used in future research as a reliable clustering assessment tool for fuzzy and probabilistic clustering models beyond FCM, capable of distinguishing solutions with varying levels of certainty.

Bibliography

[1] W.K. Kroeze, D.J. Sheffler and B.L. Roth*, G-protein-coupled receptors at a glance*, Science **116**, 4867-4869, (2003)

[2] H. Bräuner-Osborne, P. Wellendorph and A. A. Jensen, *Structure pharmacology and therapeutic prospects of family C G-Protein Coupled Receptors*, Current Drug Targets, **8**, 169-184, (2007).

[3] M.C. Lagerström and H.B. Schiöth, *Structural diversity of G protein-coupled receptors and significance for drug discovery*, Nature Reviews Drug Discovery **7**, 339-358 (2008).

[4] J. Nathans and D.S. Hogness, *Isolation, sequence analysis, and intron–exon arrangement of the gene encoding bovine rhodopsin*. Cell **34**, 807–814, (1983)

[5] P.A. Hargrave, et al., *The structure of bovine rhodopsin*, Biophysics of Structure and Mechanism, **9**, 235–244 (1983)

[6] A. Marasani, V. Talla, J.R. Gottemukkala, D. Rudrapati, *Cell signaling: Role of GPCRS*, Archives of Applied Science Research, **2**(5):363-377 (2010)

[7] T.K. Attwood and J. B. Findlay, *Design of a discriminating fingerprint for G ‐protein ‐coupled receptors*, Protein Engineering, Design and Selection, **6**, 167–176 (1993)

[8] E.S. Lander, et al*., Initial sequencing and analysis of the human genome*, Nature **409**, 860–921 (2001)

[9] J.C. Venter, et al., *The sequence of the human genome*, Science **291**, 1304–1351 (2001)

[10] T.K. Bjarnadóttir, D.E. Gloriam, S.H. Hellstrand, H. Kristiansson, R. Fredriksson, H.B. Schiöth, *Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse*, Genomics **88**(3), 263–73 (2006)

[11] X. Ding, X. Zhao and A. Watts, *G-protein-coupled receptor structure, ligand binding and activation as studied by solid-state NMR spectroscopy*, Biochemical Journal, **450**, 443-457 (2013)

[12] D.K. Vassilatis, J.G. Hohmann, H. Zeng, F. Li,J.E. Ranchalis, M.T. Mortrud, A. Brown, S.S. Rodriguez, J.R. Weller, A.C. Wright, J.E. Bergmann, G.A. Gaitanaris., *The G protein-coupled receptor repertoires of human and mouse*, Proceedings of the National Academy of Sciences of the United States of America, **100**(8): 4903-4908 (2003)

[13] A.J. Trewavas and R. Malho, *Signal perception and transduction: The origin of the phenotype,* Plant Cell, **9**, 1181–1195, (1997)

[14] F. Hucho and K. Buchner*, Signal transduction and protein kinases: the long way from the plasma membrane into the nucleus*, Naturwissenschaften, **84**, 281–290, (1997)

[15] N. Tuteja, *Signaling through G protein coupled receptors,* Plant Signaling and Behavior, **4**(10): 942-947 (2009)

[16] B.K. Kobilka*, G protein Coupled Receptors structure and activation*, Biochemical and Biophysical Acta, **1768**(4): 794-807 (2007)

[17] O. Civelli, R.K. Reinscheid, Y. Zhang, Z. Wang, R. Fredriksson, H.B. Schiöth, *G protein-coupled receptor deorphanizations*, Annual Review of Pharmacology and Toxicology, **53**,127-146 (2013)

[18] L. Chun, W. Zhang and J. Liu, *Structure and ligand recognition of class C GPCRs*, Acta Pharmacologica Sinica, **33**: 312–323 (2012)

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

[20] P. Rondard, C. Goudet, J. Kniazeff, J.P. Pin and L. Prezeau, *The complexity of their activation mechanism opens new possibilities for the modulation of mGlu and GABAB class C G protein-coupled receptors,* Neuropharmacology, **60**, 82–92 (2011)

[21] S. Urwyler, *Allosteric modulation of family C G-protein-coupled receptors: from molecular insights to therapeutic perspectives*, Pharmacological Reviews, **63**: 59–126 (2011)

[22] H. Bräuner-Osborne, P. Wellendorph and A. A. Jensen, *Structure, pharmacology and therapeutic prospects of family C G-Protein Coupled Receptors*, Current Drug Targets, **8**, 169-184 (2007)

[23] C.M. Niswender and P.J. Conn, *Metabotropic Glutamate Receptors: Physiology, pharmacology, and disease*, Annual Review of Pharmacology and Toxicology, **50**, 295-322, (2010)

[24] K.A. Johnson, P.J. Conn and C.M. Niswender, *Glutamate receptors as therapeutic targets for Parkinson's disease*, CNS & Neurological Disorders - Drug Targets, **8**(6):475-491 (2009)

[25] G. Dolen, R.L. Carpenter, T.D. Ocain and M.F. Bear, *Mechanism-based approaches to treating fragile X*, Pharmacology & Therapeutics, **127**(1):78-93 (2010)

[26] M.J. Marino, D.L. Williams Jr, J.A. O'Brien, O. Valenti, T.P. McDonald, M.K. Clements, *et al*., *Allosteric modulation of group III metabotropic glutamate receptor 4: a potential approach to Parkinson's disease treatment*, Proceedings of the National Academy of Sciences of the United States of America, **100**, 13668-13673 (2003)

[27] P.J. Conn, C.W. Lindsley and C.K. Jones, *Activation of metabotropic glutamate receptors as a novel approach for the treatment of schizophrenia*, Trends in Pharmacological Sciences, **30**: 25–31 (2009)

[28] G. Nelson, M.A. Hoon, J. Chandrashekar, Y. Zhang, N.J. Ryba and C.S. Zuker, *Mammalian sweet taste receptors*. Cell, **106**: 381–90 (2001)

[29] N. Kunishima, Y. Shimad, Y. Tsuji, T. Sato, M. Yamamoto, T. Kumasaka, et al., *Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor*, Nature, **407**: 971–7 (2000)

[30] S. Yin and C.M. Niswender, *Progress toward advanced understanding of metabotropic glutamate receptors: structure, signaling and therapeutic indications*, Cellular Signaling, **26**(10) 2284–2297 (2014)

[31] V. Katritch, V. Cherezov, and R.C. Stevens. *Structure-function of the G protein coupled receptor superfamily*, Annual Review of Pharmacology and Toxicology, **53**(1):531-556, (2013)

[32] 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, **42**(D1):D422-D425 (2014)

[33] K. Palczewski, T. Kumasaka, T. Hori, C.A. Behnke, H. Motoshima, B.A. Fox, I. Le Trong, D.C. Teller, T. Okada, R.E. Stenkamp, M. Yamamoto and M. Miyano, *Crystal structure of rhodopsin: A G protein-coupled receptor*, Science, **289**(5480):739-745 (2000)

[34] V. Katritch, V. Cherezov, and R. C. Stevens, *Structure-function of the G protein-coupled receptor superfamily,* Annual Review of Pharmacology and Toxicology, **53**, 531–556 (2013)

[35] D. Wacker, C. Wang, V. Katritch, G. W. Han, X. P. Huang, E. Vardy, J. D. McCorvy, Y. Jiang, M. Chu, F. Y. Siu, W. Liu, H. E. Xu, V. Cherezov, B. L. Roth and R. C. Stevens, *Structural features for functional selectivity at serotonin receptors,* Science **340**: 615-619, (2013)

[36] C. Wang, Y. Jiang, J. Ma, H. Wu, D. Wacker, V. Katritch, G. W. Han, W. Liu, X. P. Huang, E. Vardy, J. D. McCorvy, X. Gao, X. E. Zhou, K. Melcher, C. Zhang, F. Bai, H. Yang, L. Yang, H. Jiang, B. L. Roth, V. Cherezov, R. C. Stevens and H. E. Xu, *Structural basis for molecular recognition at serotonin receptors,* Science, **340**: 610-614, (2013)

[37] C. Wang, H. Wu, V. Katritch, G. W. Han, X. P. Huang, W. Liu, F. Y. Siu, B. L. Roth, V. Cherezov and R. C. Stevens*, Structure of the human smoothened receptor bound to an antitumour agent*, Nature, **497**: 338-343, (2013)

[38] K. Hollenstein, J. Kean, A. Bortolato, R.K. Cheng, A.S. Dore, A. Jazayeri, R.M. Cooke, M. Weir, F.H. Marshall, *Structure of class B GPCR corticotropin-releasing factor receptor 1*, Nature, **499**, 438-443, (2013)

[39] F.Y. Siu, M. He, C. de Graaf, G. W. Han, D. Yang, Z. Zhang, C. Zhou, Q. Xu, D. Wacker, J. S. Joseph, W. Liu, J. Lau, V. Cherezov, V. Katritch, M. W. Wang and R. C. Stevens, *Structure of the human glucagon class B G-protein-coupled receptor*, Nature, **499**: 444-449, (2013)

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

[41] A. Gutcaits, P. Prusis, C. Post, T. Lundstedt, J.E.S. Wikberg, M. Lapinsh, *Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences*. Protein Science, **11**(4):795-805, (2002)

[42] 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, **41**(14):2481-2491, (1998)

[43] C. Caragea, A. Silvescu, P. Mitra, *Protein sequence classification using feature hashing*, Proteome Science, **10.Suppl**, S14, (2012)

[44] R. Cruz-Barbosa, A. Vellido, and J. Giraldo, *Advances in semi-supervised alignment-free classication of G protein-coupled receptors*, In Proceeding of the International Work-

Conference on Bioinformatics and Biomedical Engineering (IWBBIO'13), pp. 759-766, (2013)

[45] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum, New York, 1981.

[46] J.C. Bezdek, R. Ehrlich, W. Full, *FCM: The Fuzzy c-Means clustering algorithm*, Computers & Geosciences, **10**(2-3), 191-203, (1984)

[47] M. Julio-Pieper, P.J. Flor, T.G. Dinan and J.F. Cryan, *Exciting times beyond the brain: Metabotropic Glutamate Receptors in peripheral and non-neural tissues*, Pharmacological Reviews, **63**(1), 35-58 (2011)

[50] H. Ohashi , T. Maruyama , H. Higashi-Matsumoto, T. Nomoto, S. Nishimura, Y. Takeuchi, *A novel binding assay for metabotropic glutamate receptors using [3H] L-quisqualic acid and recombinant receptors*, Zeitschrift fur Naturforschung C, **57**(3-4):348-355 (2002)

[51] E. Hinoi, K. Ogita, Y. Takeuchi, H. Ohashi, T. Maruyama and Y. Yoneda , *Characterization with [3H] quisqualate of group I metabotropic glutamate receptor subtype in rat central and peripheral excitable tissues*, Neurochemistry International, **38**(3):277-285 (2001)

[52] Z. Chua and J.J. Hablit, *Quisqualate induces an inward current via mGluR activation in neocortical pyramidal neurons*, Brain Research, **879**(1–2), 88–92 (2000)

[53] S.R. Platt, *The role of glutamate in central nervous system health and disease – A review*, The Veterinary Journal, **173**(2), 278–286 (2007)

[54] D. Dembele, P. Kastner, *Fuzzy C-means method for clustering microarray data*, Bionformatics, **19**(8), 973- 980 (2003)

[55] M.I. Cárdenas, A. Vellido, C. König, R. Alquézar and J. Giraldo, *Exploratory visualization of misclassified GPCRs from their transformed unaligned sequences using manifold learning techniques,* In F. Ortuño, I. Rojas (eds.): Procs. of the 2nd International Work-Conference on Bioinformatics and Biomedical Engineering, pp.623-630, (2014)

[56] M.I. Cárdenas, A. Vellido, J. Giraldo, *Exploratory visualization of Metabotropic Glutamate Receptor subgroups through manifold learning*, 17th International Conference of the Catalan Association of Artificial Intelligence, CCIA, (2014)

[57] C. König, R. Cruz-Barbosa, R. Alquézar and A. Vellido, A, *SVM-Based Classification of Class C GPCRs from Alignment-Free Physicochemical Transformations of Their Sequences,* 2nd International Workshop on Pattern Recognition in Proteomics, Structural Biology and Bioinformatics, (PR PS BB) 2013, 17th International Conference on Image Analysis and Processing (ICIAP) In A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 336–343, Springer (2013)

[58] C. König, A. Vellido, R. Alquézar and J. Giraldo, *Misclassification of class C G-protein-coupled receptors as a label noise problem*, In Proceedings of the 22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2014), Bruges, Belgium. pp. 695-700 (2014)

[59] Gprotein-coupled receptors, Wikipedia entry. URL: http://en.wikipedia.org/wiki/G_protein-coupled_receptor. Last accessed 28/08/2014

[60] P.J. Conn, A. Christopoulos and C.W. Lindsley, *Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders*, Nature Reviews Drug Discovery, **8**, 41-54, (2009)

[61] B. Snyder, *Where are the new drugs? The push to improve the pipeline*, Lens, a Publication of Vanderbit Medical Center, (2005)

[62] X.L. Xie and G. Beni, *A validity measure for fuzzy clustering*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **13**(8), 841-847, (1991)

[63] A.K. Jian, Data clustering: 50 years beyond K-means. Pattern Recognition Letters, **31**(8), 651-666, (2010)

[64] B. Frenay, M. Verleysen, *Classification in the presence of label noise: A survey*, IEEE Transactions on Neural Networks and Learning Systems, **25**(5), 845-869, (2014)

[65] P.J.G. Lisboa, T.A. Etchells, I.H. Jarman and S.J. Chambers, Finding reproducible cluster partitions for the K-Means algorithm. BMC Bioinformatics, 14(Suppl 1), S8, (2013)

Appendix A

Supplementary materials to Chapter 5: Experimental Study

Concerning sub-section 5.1: Experiments with a Fixed Number of Clusters; 5.1.1 Fuzzy c-Means for class C GPCRs

Class C GPCR: Detailed numerical results corresponding to the graphics in Figs. 5.1 to 5.3, for the different data transformations.

- **AAC Transformation**

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7
class 1	59.18%	5.02%	54.00%	1.55%	47.76%	3.04%	14.19%
class 2	3.67%	13.81%	1.50%	0.00%	0.00%	1.14%	0.00%
class 3	31.43%	0.42%	40.50%	0.00%	11.94%	1.52%	12.21%
class 4	1.22%	43.10%	1.50%	51.30%	2.99%	21.67%	25.41%
class 5	4.08%	17.99%	2.00%	43.52%	7.46%	41.83%	44.88%
class 6	0.41%	17.99%	0.00%	0.52%	2.99%	17.11%	3.30%
class 7	0.00%	1.67%	0.50%	3.11%	26.87%	13.69%	0.00%

Table A.1: Class specificity in each cluster of CGPCR_AAC dataset.

- **AAC Transformation**

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7
class 1	98.13%	3.38%	4.95%	2.95%	84.42%	3.58%	15.77%
class 2	0.00%	0.48%	0.50%	0.00%	7.54%	11.11%	0.00%
class 3	0.00%	0.00%	94.55%	0.00%	4.02%	0.72%	2.51%
class 4	0.00%	24.15%	0.00%	57.33%	1.00%	19.35%	36.56%
class 5	1.87%	62.80%	0.00%	39.24%	3.02%	53.41%	4.30%
class 6	0.00%	5.31%	0.00%	0.42%	0.00%	7.53%	24.73%
class 7	0.00%	3.86%	0.00%	0.00%	0.00%	4.30%	16.13%

Table A.2: Class specificity in each cluster of CGPCR_ACC dataset

- **Digram Transformation**

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7
class 1	98.21 %	13.10%	4.00%	0.00%	96.65%	1.46%	4.91%
class 2	0.00%	11.23%	0.50%	0.00%	1.68%	0.98%	0.00%
class 3	0.00%	4.01%	94.50%	0.00%	0.56%	0.00%	1.84%
class 4	0.00%	36.63%	0.50%	61.73%	0.00%	10.73%	7.98%
class 5	1.79%	23.80%	0.50%	65.74%	1.12%	380.5%	74.23%
class 6	0.00%	8.56%	0.00%	2.53%	0.00%	21.95%	11.04%
class 7	0.00%	2.67%	0.00%	0.00%	0.00%	26.83%	0.00%

Table A.3: Class specificity in each cluster of CGPCR_Digram dataset

mGluR: Detailed numerical results corresponding to the graphics in Figs. 5.4 to 5.6, for the different data transformations.

- **AAC Transformation**

	cluster 1	cluster 2	cluster 3
class 1	0.00%	50.00%	9.52%
class 2	50.00%	4.08%	30.95%
class 3	50.00%	45.92%	59.52%

Table A.4: Class specificity in each cluster of mGluR_AAC dataset

- **ACC Transformation**

	cluster 1	cluster 2	cluster 3
class 1	74.68%	2.78%	0.00%
class 2	1.27%	75.00%	3.81%
class 3	24.05%	22.22%	96.19%

Table A.5: Class specificity in each cluster of mGluR _ACC dataset

- **Digram Transformation**

	cluster 1	cluster 2	cluster 3
class 1	98.33%	2.11%	0.00%
class 2	1.67%	56.84%	3.96%
class 3	0.00%	41.05%	96.04%

Table A.6: Class specificity in each cluster of mGluR _Digram dataset

Concerning sub-section 5.1: Experiments with a Fixed Number of Clusters;
5.1.2 K-means for class C GPCRs

Class C GPCR: Detailed numerical results corresponding to the graphics in Figs. 5.7 to 5.9, for the different data transformations.

- **AAC Transformation**

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7
class 1	54.81%	2.71%	56.12%	1.06%	61.11%	0%	13.86%
class 2	8.52%	5.67%	1.02%	0%	0%	0%	0%
class 3	30%	0.25%	37.76%	0%	29.63%	0%	10.62%
class 4	1.11%	33.25%	2.04%	51.85%	1.85%	23.21%	26.55%
class 5	4.44%	34.48%	3.06%	43.92%	7.41%	5.36%	42.48%
class 6	0.74%	15.52%	0%	0%	0%	26.79%	6.49%
class 7	0.37%	8.13%	0%	3.17%	0%	44.64%	0%

Table A.7: Class specificity in each cluster of CGPCR_AAC dataset.

- **AAC Transformation**

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7
class 1	96.54%	0.00%	2.67%	0.00%	51.52%	1.06%	3.02%
class 2	0.00%	0.00%	0.67%	0.00%	1.21%	11.87%	0.00%
class 3	0.38%	0.00%	96.67%	0.00%	37.58%	0.00%	0.00%
class 4	0.77%	5.15%	0.00%	52.66%	0.61%	38.79%	37.93%
class 5	2.31%	87.50%	0.00%	46.28%	4.85%	40.63%	7.76%
class 6	0.00%	5.88%	0.00%	1.06%	0.61%	7.39%	27.16%
class 7	0.00%	1.47%	0.00%	0.00%	3.64%	0.26%	24.14%

Table A.8: Class specificity in each cluster of CGPCR_ACC dataset

- **Digram Transformation**

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7
class 1	98.65%	1.01%	0%	0%	44.57%	0%	44.23%
class 2	0%	0%	0%	0%	25%	1.02%	0%
class 3	0%	0%	100%	0%	22.83%	0%	43.27%
class 4	0%	35.93%	0%	61.97%	1.09%	11.17%	0.96%
class 5	1.35%	51.26%	0%	35.21%	4.89%	37.56%	1.92%
class 6	0%	11.56%	0%	2.82%	0.54%	22.34%	2.88%
class 7	0%	0.25%	0%	0%	1.09%	27.92%	6.73%

Table A.9: Class specificity in each cluster of CGPCR_Digram dataset

mGluR: Detailed numerical results corresponding to the graphics in Figs. 5.10 to 5.12, for the different data transformations.

- **AAC Transformation**

	cluster 1	cluster 2	cluster 3
class 1	52.69%	0%	9.16%
class 2	3.23%	50%	30.53%
class 3	44.09%	50%	60.31%

Table A.10: Class specificity in each cluster of mGluR_AAC dataset

- **ACC Transformation**

	cluster 1	cluster 2	cluster 3
class 1	98.21%	6.19%	0%
class 2	1.79%	55.67%	3.88%
class 3	0%	38.14%	96.12%

Table A.11: Class specificity in each cluster of mGluR _ACC dataset

- **Digram Transformation**

	cluster 1	cluster 2	cluster 3
class 1	98.18%	5.15%	0%
class 2	1.82%	39.71%	6.15%
class 3	0%	55.15%	93.85%

Table A.12: Class specificity in each cluster of mGluR _Digram dataset

Appendix B

Supplementary materials to Chapter 5: Experimental Study

Concerning sub-section
5.1.1 Fuzzy c-Means for class C GPCRs

Some extra experimental results that were obtained using the FCM algorithm are presented in this appendix. They concern the analysis of the number of cases that were assigned to a cluster with membership values between specific value ranges was calculated and displayed as histograms, which are reported in section 5.2.2. This will inform us to be the level of certainty with which the algorithm is assigning sequences to clusters (higher certainty being an indication of crisp decisions and thus clearly separated clusters). Furthermore, experiments were carried out with different threshold values. If the highest membership value of a case was greater or equal to the threshold, this case was assigned to the corresponding cluster, otherwise the case was rejected (an *abstentionist* system that only commits when a certain degree of decision certainty is achieved). The relation between the accuracy and the threshold values and the number of the rejected cases and the threshold values is displayed in several figures in section 5.2.3.

The tables with the numeric values of the variables used to create each of the figures in the following sub-sections are also provided as supplementary material in this Appendix.

5.1.1.1 Membership Values range

For each dataset, the number of the data points with maximum membership values between 1 and 0.9, 0.9 and 0.8, 0.8 and 0.7, 0.7 and 0.6, 0.6 and 0.5, 0.5 and 0.4, 0.4 and 0.3, 0.3 and 0.2, 0.2 and 0.1, and finally 0.1 and 0 was calculated and displayed in the following figures.

CGPCR

CGPCR_AAC

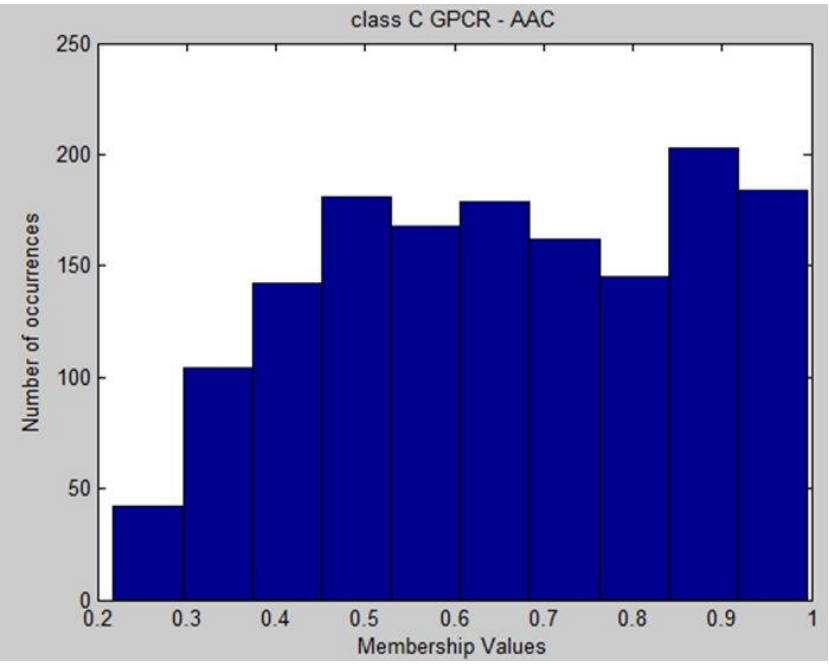

Figure B.1: CGPCR_AAC dataset's histogram.

CGPCR_ACC

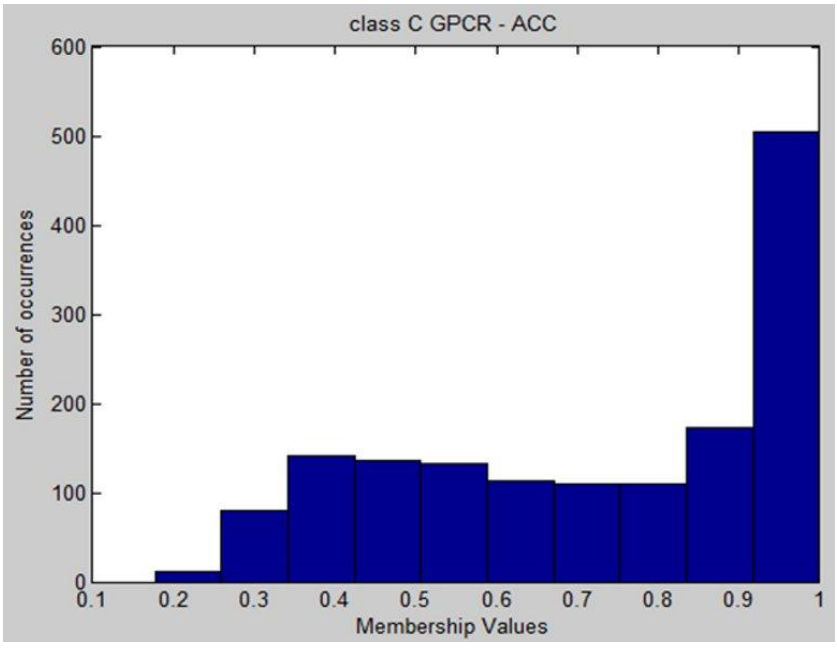

Figure B.2: CGPCR_ACC dataset's histogram.

CGPCR_Digram

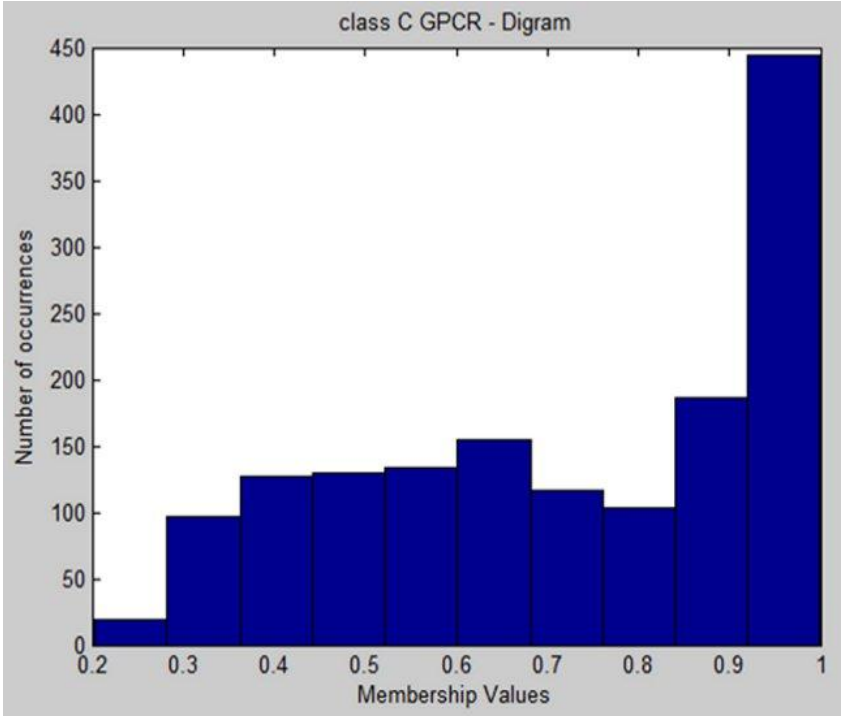

Figure B.3: CGPCR_Digram dataset's histogram.

mGluR

mGluR_AAC

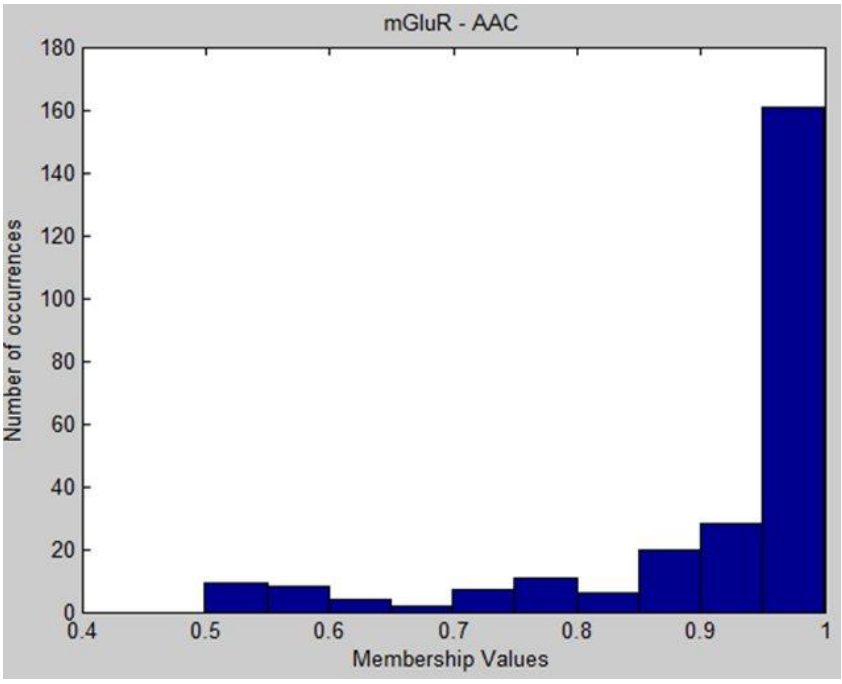

Figure B.4: mGluR_AAC dataset's histogram.

mGluR_ACC

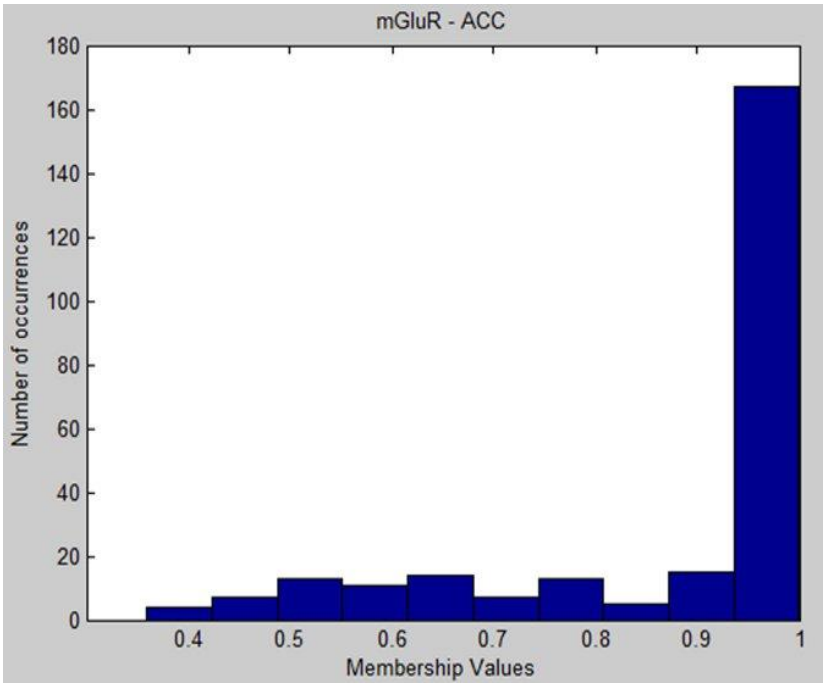

Figure B.5: mGluR_ACC dataset's histogram

mGluR_Digram

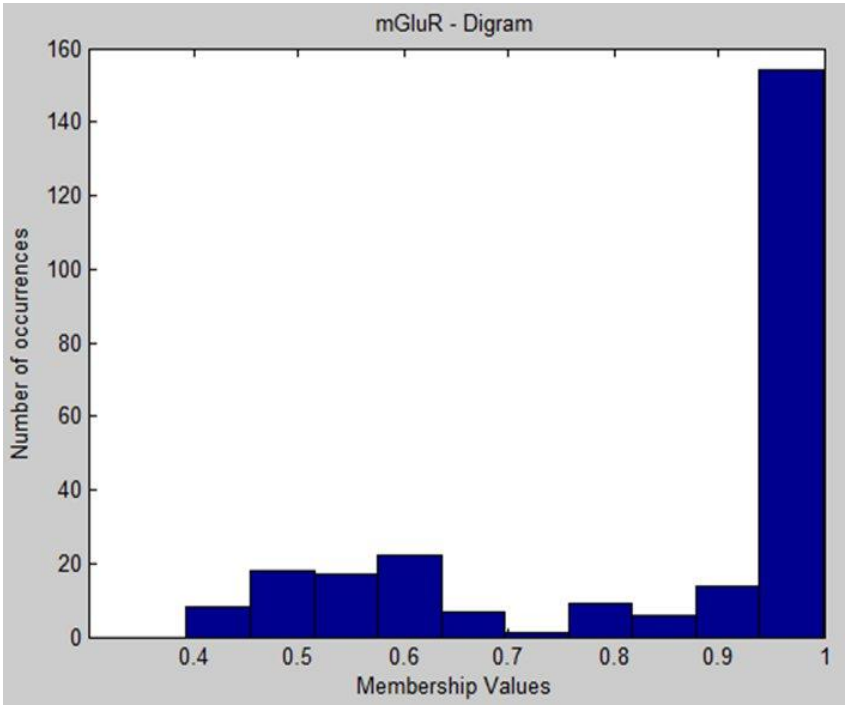

Figure B.6: mGluR_Digram dataset's histogram

One of the obvious advantages of using a fuzzy model such as FCM is that it can estimate the degree of cluster membership for each GPCR sequence. Even if we ultimately make a *crisp* cluster assignment decision on the basis of the maximum degree of membership across clusters, we can still have information about the level of certainty with which that *crisp* assignment is made.

In that sense, Figures B.1 to B.6 are very informative and, importantly, rather consistent with the results discussed in the previous section.

For the complete class C GPCR dataset (B.1 to B.3), the level of uncertainty yielded by the model based on the AAC transformed data is quite high, with a large number of sequences showing maximum membership levels at around values of 0.5 or 0.6; that is, cluster assignments with not too-high levels of confidence. In comparison, the models based on the ACC and Digram (specially the former) data transformations concentrate the membership degree values in the top decile, which means that their cluster assignments are extremely certain.

The corresponding results for the mGluR subtype reported in Figures B.4 to B.6 partially differ from the previous in the sense that, although all data transformations yield a very certain cluster memberships, the simple AAC transformation seems to be the most certain of them all. Interestingly, this means that, despite the certainty of the model, the information conveyed by this transformation does not conform to the standard data partition of mGluR into three groups.

5.1.1.2 Membership threshold analysis for an abstaining system

For some problems, it is important to reach a decision stage for each of the data items analyzed. For some other types of problems, though, reaching a decision for each case may be less relevant than figuring out which cases are worth reaching a decision at all costs. In diagnostic decision making, for instance, it might not be worth (or adequate) deciding on a given case diagnosis unless this decision is reached with sufficient certainty. Sometimes, a diagnosis made on the basis of weak evidence can be far more damaging

than abstaining from making a diagnosis.

In the fields of Machine Learning and Computational Intelligence, if the model reaching a decision is a supervised classifier, this type of approaches is known as *abstaining classifiers*.

In the type of problem addressed by the current thesis, we might speculate with the possibility of not making a decision about cluster membership unless the certainty of such assignment reaches a threshold level.

Figures B.7 to B.12 are a summary of the results obtained when applying this *abstaining* approach. We get two graphs in each: the first reflecting the evolution of the accuracy as the threshold shifts, and the second complementing this with information on how many cases are *reject for decision* (that is, in how many cases the FCM decision maker refrains from deciding, or *abstains*) as the threshold shifts. Note that if the highest membership value for a given transformed sequence is lower than the threshold, then the case is rejected and thus not counted in the calculation of the overall accuracy of the FCM algorithm. In other words, the resulting accuracy is calculated as the ratio of correctly assigned non-rejected cases to the total of non-rejected cases and it should only be considered as an indicator of model certainty levels. As the rejected cases are not included in the accuracy calculation, such accuracy is prone to increase as the threshold is increased.

Figures B.7 to B.9 for the class C GPCR complete data set reflect that the starting point accuracies (with no rejection) are already quite low in all models, especially for the AAC transformation (consistent with the results in Figures 5.1 to 5.3), and that there is a considerable increase of rejections in all models from a threshold of approximately 0.3. The number of cases rejected is overall higher for the simple AAC transformation. The evolution of the accuracy is not monotonous as the threshold increases, which is consistent with the results shown in Figures B.1 to B.3.

The corresponding Figures B.10 to B.12 for the mGluR subtype results reflect a similar situation, with a few differences: the initial accuracies are homogeneously higher (over 80% for ACC and Digram, but just over 50% for AAC) and raise even over 90% for Digram

in the last decile. The bulk of the rejection does not start in earnest until the 0.4 threshold is not reached. Note that the highest level of rejected cases for all transformations is a much lower percentage of all cases than that obtained for the complete C GPCR data set. Again, these results are consistent with those reported in Figures 5.4 to 5.6 and B.4 to B.6.

CGPCR

CGPCR_AAC

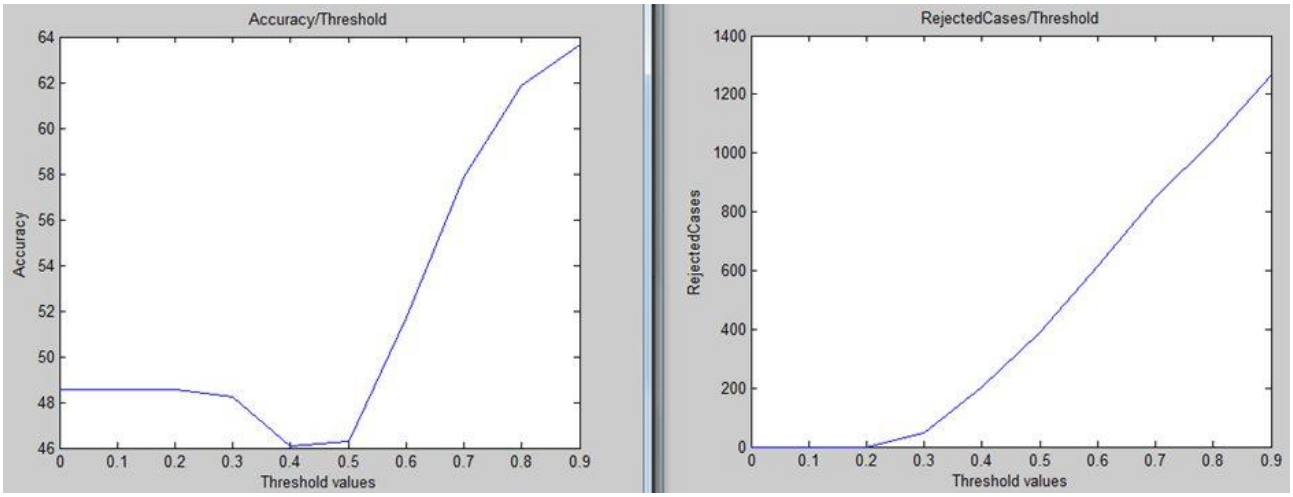

Figure B.7: The Accuracy and the number of rejected cases compared with the threshold values for CGPCR_AAC dataset

CGPCR_ACC

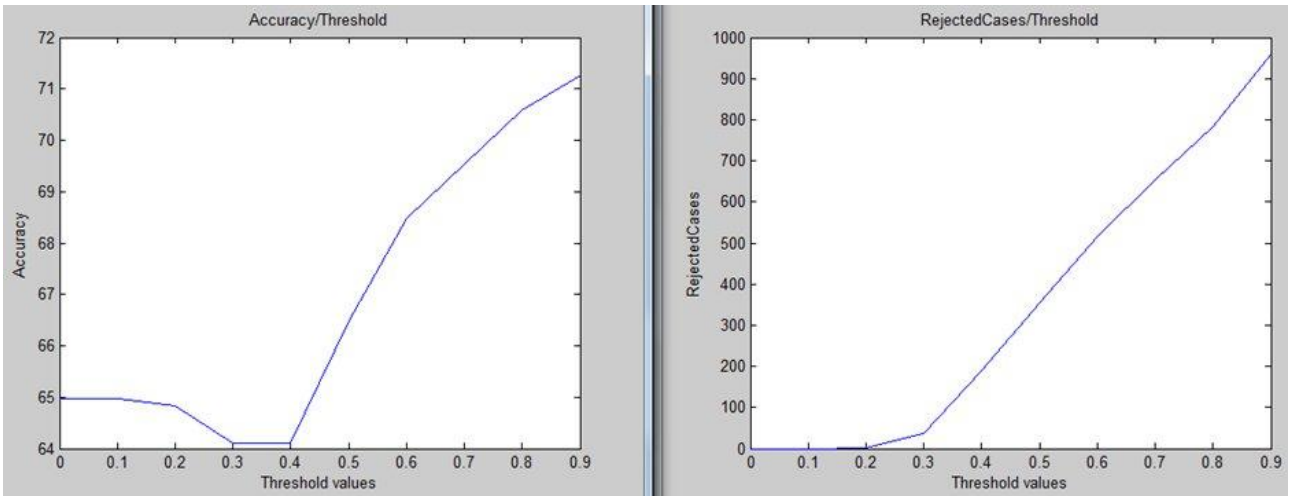

Figure B.8: The Accuracy and the number of rejected cases compared with the threshold values for CGPCR_ACC dataset

CGPCR_Digram

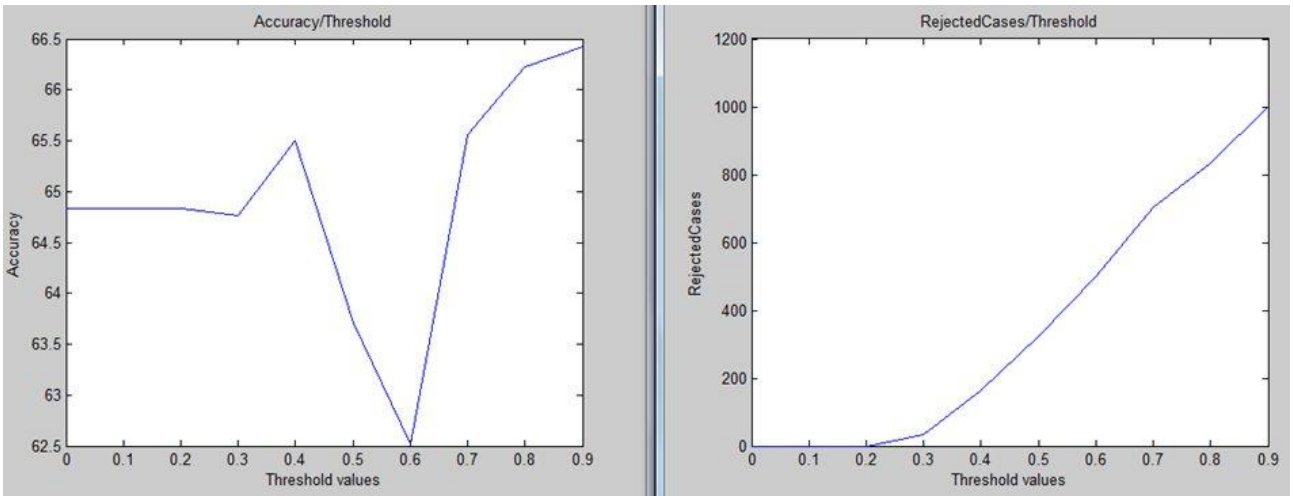

Figure B.9: The Accuracy and the number of rejected cases compared with the threshold values for CGPCR_Digram dataset

mGluR

mGluR_AAC

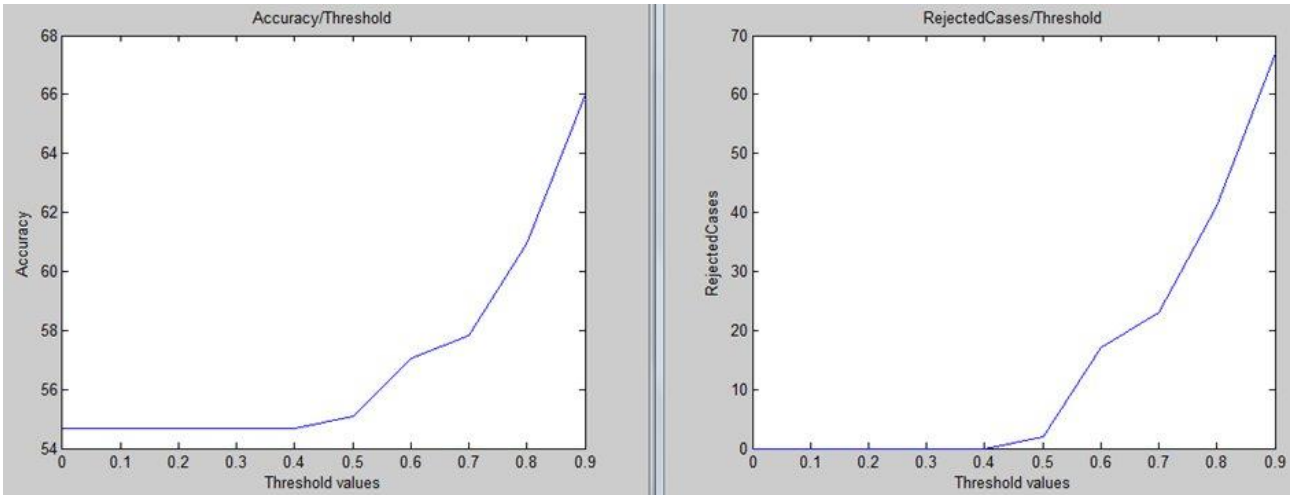

Figure B.10: The Accuracy and the number of rejected cases compared with the threshold values for mGluR_AAC dataset

mGluR_ACC

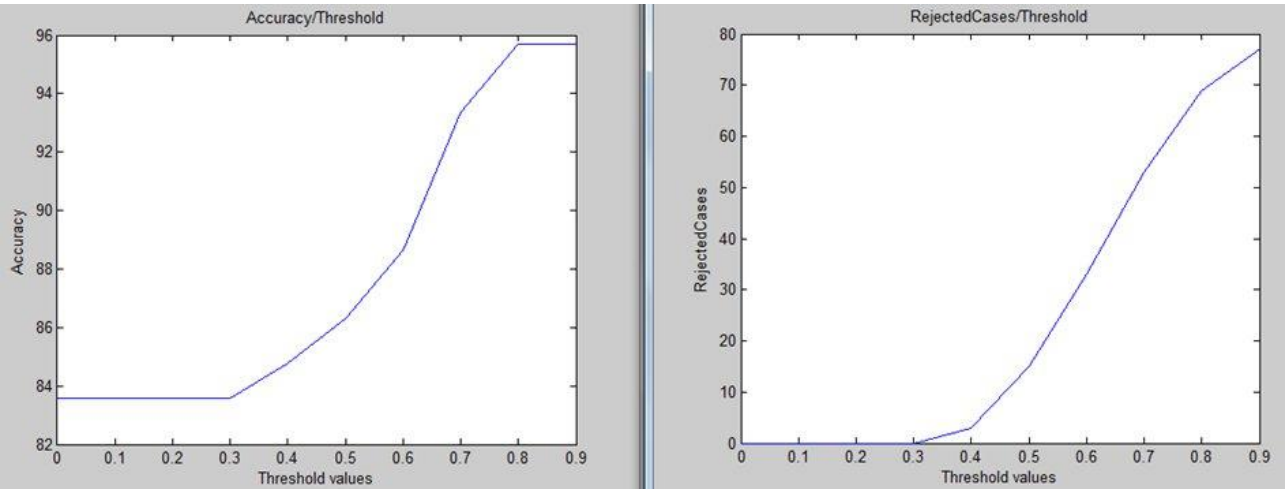

Figure B.11: The Accuracy and the number of rejected cases compared with the threshold values for mGluR_ACC dataset

mGluR_Digram

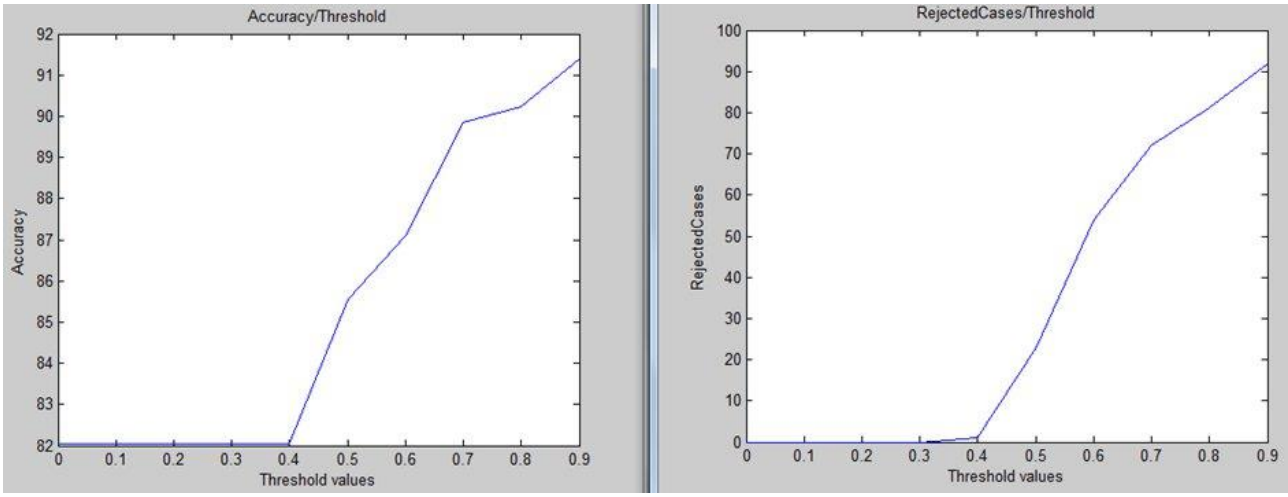

Figure B.12: The Accuracy and the number of rejected cases compared with the threshold values for mGluR_Digram dataset

Supplementary tables including the numerical values corresponding to the Figures included in this Appendix B.

5.1.1.1 Membership Values range*: Figures B.1 to B.3*

CGPCR_AAC

Membership Values	Number of sequences
0.0 – 0.1	0
0.1 – 0.2	0
0.2 – 0.3	33
0.3 – 0.4	129
0.4 – 0.5	164
0.5 – 0.6	175
0.6 – 0.7	203
0.7 – 0.8	129
0.8 – 0.9	168
0.9 - 1	509

Table B.13: Membership Values for the CGPCR_AAC dataset

CGPCR_ACC

Membership Values	Number of sequences
0.0 – 0.1	0
0.1 – 0.2	2
0.2 – 0.3	36
0.3 – 0.4	152
0.4 – 0.5	166
0.5 – 0.6	161
0.6 – 0.7	137
0.7 – 0.8	129
0.8 – 0.9	177
0.9 - 1	550

Table B.14: Membership Values for the CGPCR_ACC dataset

CGPCR_Digram

Membership Values	Number of sequences
0.0 – 0.1	0
0.1 – 0.2	0
0.2 – 0.3	33
0.3 – 0.4	129
0.4 – 0.5	164
0.5 – 0.6	175
0.6 – 0.7	203
0.7 – 0.8	129
0.8 – 0.9	168
0.9 - 1	509

Table B.15: Membership Values for the CGPCR_Digram dataset

Membership Values range*: Figures B.4 to B.6*

mGluR_AAC

Membership Values	Number of sequences
0.0 – 0.1	0
0.1 – 0.2	0
0.2 – 0.3	0
0.3 – 0.4	0
0.4 – 0.5	0
0.5 – 0.6	7
0.6 – 0.7	5
0.7 – 0.8	6
0.8 – 0.9	7
0.9 - 1	231

Table B.16: Membership Values for the mGluR_AAC dataset

mGluR _ ACC

Membership Values	Number of sequences
0.0 – 0.1	0
0.1 – 0.2	0
0.2 – 0.3	0
0.3 – 0.4	0
0.4 – 0.5	2
0.5 – 0.6	7
0.6 – 0.7	7
0.7 – 0.8	8
0.8 – 0.9	16
0.9 - 1	216

Table B.17: Membership Values for the mGluR _ACC dataset

mGluR _ Digram

Membership Values	Number of sequences
0.0 – 0.1	0
0.1 – 0.2	0
0.2 – 0.3	0
0.3 – 0.4	7
0.4 – 0.5	51
0.5 – 0.6	13
0.6 – 0.7	9
0.7 – 0.8	12
0.8 – 0.9	25
0.9 - 1	139

Table B.18: Membership Values for the mGluR_Digram dataset

5.1.1.2 Threshold. Membership Values range*: Figures B.7 to B.9*

CGPCR_AAC

Threshold	Accuracy	Rejected Cases
0.0	48.54%	0
0.1	48.54%	0
0.2	48.54%	0
0.3	48.21%	46
0.4	46.09%	201
0.5	46.29%	389
0.6	51.72%	618
0.7	57.88%	849
0.8	61.85%	1042
0.9	63.64%	1269

Table B.19: The Accuracy and the number of rejected cases compared with the threshold values for CGPCR_AAC dataset

CGPCR_ACC

Threshold	Accuracy	Rejected Cases
0.0	64.96%	0
0.1	64.96%	0
0.2	64.83%	2
0.3	64.10%	38
0.4	64.10%	190
0.5	66.49%	356
0.6	68.47%	517
0.7	69.53%	654
0.8	70.59%	783
0.9	71.25%	960

Table B.20: The Accuracy and the number of rejected cases compared with the threshold values for CGPCR_ACC dataset

CGPCR_Digram

Threshold	Accuracy	Rejected Cases
0.0	64.83%	0
0.1	64.83%	0
0.2	64.83%	0
0.3	64.76%	33
0.4	65.49%	162
0.5	63.70%	326
0.6	62.51%	501
0.7	65.56%	704
0.8	66.22%	833
0.9	66.42%	1001

Table B.21: The Accuracy and the number of rejected cases compared with the threshold values for CGPCR_Digram dataset

Threshold. Membership Values range*: Figures B.10 to B.12*

mGluR_AAC

Threshold	Accuracy	Rejected Cases
0.0	54.68%	0
0.1	54.68%	0
0.2	54.68%	0
0.3	54.68%	0
0.4	54.68%	0
0.5	55.07%	2
0.6	57.03%	17
0.7	57.81%	23
0.8	60.93%	41
0.9	66.01%	67

Table B.22: The Accuracy and the number of rejected cases compared with the threshold values for mGluR_AAC dataset

mGluR _ ACC

Threshold	Accuracy	Rejected Cases
0.0	83.59%	0
0.1	83.59%	0
0.2	83.59%	0
0.3	83.59%	0
0.4	84.76%	3
0.5	86.32%	15
0.6	88.67%	33
0.7	93.35%	53
0.8	95.70%	69
0.9	95.70%	77

Table B.23: The Accuracy and the number of rejected cases compared with the threshold values for mGluR_ACC dataset

mGluR _ Digram

Threshold	Accuracy	Rejected Cases
0.0	82.03%	0
0.1	82.03%	0
0.2	82.03%	0
0.3	82.03%	0
0.4	82.03%	1
0.5	85.54%	23
0.6	87.10%	54
0.7	89.84%	72
0.8	90.23%	81
0.9	91.40%	92

Table B.24: The Accuracy and the number of rejected cases compared with the threshold values for mGluR_Digram dataset