Classification of GPCRs Using Family Specific Motifs

by Murat Can Cobanoglu

Submitted to the Graduate School of Sabancı University
in partial fulfillment of the requirements for the degree of
Master of Science

Sabanci University

Spring, 2010

Classification of GPCRs
Using Family Specific Motifs

Approved by:

Assoc.Prof.Dr. Yucel Saygin ....
(Dissertation Supervisor)

Assoc.Prof.Dr. Ugur Sezerman ....
(Dissertation Co-Supervisor)

Assoc.Prof.Dr. Erkay Savas ....

Assist.Prof.Dr. Husnu Yenigun ....

Assist.Prof.Dr. Devrim Gozuacik ....

Date of Approval: ....

Classification of GPCRs
Using Family Specific Motifs

Murat Can Cobanoglu

CS, Master's Thesis, 2010

Thesis Supervisors: Yucel Saygin, Ugur Sezerman

Abstract

The classification of G-Protein Coupled Receptor (GPCR) sequences is an important problem that arises from the need to close the gap between the large number of orphan receptors and the relatively small number of annotated receptors. Equally important is the characterization of GPCR Class A subfamilies and gaining insight into the ligand interaction since GPCR Class A encompasses a very large number of drug-targeted receptors. In this thesis, a method for Class A subfamily classification using sequence-derived motifs which characterizes the subfamilies by discovering receptor-ligand interaction sites is proposed. The motifs that best characterize a subfamily are selected by the proposed Distinguishing Power Evaluation (DPE) technique. The experiments performed on GPCR sequence databases show that the proposed method outperforms state-of-the-art classification techniques for GPCR Class A subfamily prediction. An important contribution of this thesis is to discover key receptor-ligand interaction sites which is very important for drug design.

Classification of GPCRs
Using Family Specific Motifs

Murat Can Cobanoglu

CS, Master Tezi, 2010

Thesis Supervisors: Yucel Saygin, Ugur Sezerman

Özet

G-protein ile eşleşmiş reseptörlerin (GPER) sınıflandırılması, fonksiyonu belirlenememiş ancak amino asit dizilimi belirlenmiş çok sayıdaki reseptörün fonksiyonunu tahmin edebilmeyi mümkün kılması açısından çok önemlidir. GPER proteinleri arasında A sınıfı reseptörlerin çok sayıda ilaç tarafından hedef alınıyor olması sebebiyle, A sınıfı reseptörlerin aktivasyon mekanizmalarının derinlikli şekilde anlaşılabilmesi ise ayrıca önem teşkil etmektedir. Bu tezde, reseptörlerdeki amino asit dizilimi verisinden üretilmiş motifler kullanılarak A sınıfındaki reseptör ailelerinin sınıflandırılmasını sağlayan, ürettiği motifler yoluyla da A sınıfı reseptörlerinin aktivasyon mekanizmalarına ışık tutan bir yöntem sunulmaktadır. Alt-sınıfları en iyi şekilde tanımlayan motifleri seçebilmek için Ayrıştırı Güç Değerlendirmesi tekniğini sunuyoruz. Yapılan deneyler, geliştirdiğimiz yöntemin halıhazırda bulunan GPER proteinleri A sınıfı reseptörlerinin sınıflandırması tekniklerine kıyasla daha yüksek başarı oranları yakaladığını göstermiştir. Bu tezin bir diğer katkısı da ilaç tasarımında faydalı olabilecek, reseptör aktivasyonunda rol oynayan anahtar bölgelerin bulunmasıdır.

to my loving mother, my wise father

Acknowledgements

I wish to express my gratitudes to,

- Ugur Sezerman and Yucel Saygin for their supervision.
- Thesis comittee for their participation.
- TUBITAK for their financial support.

Contents

List of Tables

List of Figures

1 Introduction

The G-Protein Coupled Receptor (GPCR) protein sequences are of very high interest to researchers in the drug design industry and in many other areas as more than 50% of modern drugs target GPCRs [3]. These receptors control pathways and mechanisms that govern many of the important functions in many different species, including humans. The GPCRs play a key role in sensing a very diverse set of signals ranging from visual to olfactory. This is because GPCRs have a primary function in establishing the sensory and regulatory connection of the cell with the outside world as they both act as receptors for outside ligands (ligands range from photons inducing sight to small peptides inducing neurological effects) and actuators for internal processes.

The ability of the GPCRs to regulate important functions is well-recognized in the drug design efforts: some pharmaceutical research companies like Norak, Arena, 7TM, Novasite, and Predix are exclusively focused on GPCR drug discovery, while most major pharmaceutical giants have GPCR-targeting drugs such as Zyprexa of Eli Lilly, Clarinex of Schering-Plough, Zantac of GlaxoSmithKline, and Zelnorm of Novartis[3].

Due to their significant role, it is very important to be able to distinguish which ligands that a specific GPCR interacts with and which parts of the sequence have a particularly important role. The nature of this signal transduction is complex and the binding of the ligand constitutes only the first step of this process [4]. Upon binding of the ligand to the receptor, certain interactions are established which trigger conformational changes in the GPCR and initiate the signal transduction process [5]. Determining the functionally important interactions between the ligand and the receptor is of paramount importance for drug design purposes. Being able to correctly identify the sites that regulate binding of a GPCR to a ligand can significantly reduce the set of potential ligands. Achieving this goal can also enable us to assess the mechanics of ligand-activation for these receptors.

On this pursuit, sequence remains to be the primary source of information for a large number of GPCR receptors because it is extremely difficult to get the structure of these proteins with methods like X-Ray Crystallography and Nuclear Magnetic Resonance (NMR) as these methods fail to work properly on proteins that are embedded in the cell membrane. Consequently, researchers use high-throughput screening methods to discover the activating small structures that have been chemically synthesized. The aim of these screening efforts is to identify the important characteristics of the receptor. If a computational method can identify the sites that are significant in the physiology of the receptor, these high-cost screening efforts can be avoided – saving both time and resources.

As a result, there are two presiding goals for computational methods in GPCR research: firstly, to classify GPCR sequences with respect to subfamilies within Family A which contains more than 80% of human GPCRs (as shown in Figure 2), secondly and most importantly to identify the key ligand-interacting sites using the sequence alone.

In response to the above requirements, a classification technique that also pinpoints ligand-receptor interaction sites has been developed. To the best of the author's knowledge, this is the first classification technique that makes the ligand-receptor interaction sites transparent to drug designers.

The proposed technique involves identifying the frequent residue triplets in the sequence, calculating their distinguishing power among the subfamilies and deducing rules from this information. Since these triplets are specific to a subfamily where the GPCR is exposed to the ligand they should be involved in either recruiting the ligand to the receptor or actually binding the ligand. Therefore, these potential interaction sites are called key sites throughout this thesis. These rules are then used in classification and the combination of rules for a particular subfamily directs us towards the interaction sites. To be able to increase the classification quality, the locations that each of these triplets occur very frequently have been determined through statistical measures and then this information has been used to confine the classification dataset attributes to these motifs. The proposed methods have been implemented and tested on real GPCR sequences and the experiments show that the proposed methods outperform state-of-the-art classification methods. The best performing GPCR classifier classifies the GPCRpred class A dataset with 76% accuracy; whereas the proposed method demonstrates up to 90.7% accuracy on the same testing set.

Given the high performance of the proposed method, it is only natural to think that the discovered motifs pinpoint the most important binding sites. The rationale is that if these sites were not related to binding, then they would not have been conserved in all the sequences in a subfamily. However if these sites regulated binding to all the GPCR Class A ligands, then they would occur in all the sequences and they would not have any distinguishing power. The motifs that occur in all or most of the members of a particular subfamily and do not occur in other subfamilies are identified, and this hints that the proposed method identifies binding sites specific to the ligands of a particular subfamily.

2 Related Work and Contribution

A technique commonly employed in classification is Support Vector Machines (SVM) which have also been utilized in GPCR classification. One good example in which SVMs are used for GPCR classification is the GPCRpred server [6]. In GPCRpred, 20 different SVMs are built for different levels of classification where the feature vectors are derived from the dipeptide composition of each protein. The reported classification accuracy for each level of classification is quite high, ranging over 90%. Other studies indicate that SVM classification gives better results compared to BLAST and profile HMMs [7]. Despite the strong results achieved by using SVM as reported in [8, 7] SVM-based classification techniques fail to pinpoint precisely which physico-chemical properties of the receptor were decisive in determining the corresponding ligand. It would be helpful to report the common physico-chemical qualities that are attributed to a particular ligand's receptors because such information could potentially be used in drug design efforts.

Hidden Markov Models (HMM) are one of the classification tools employed in GPCR classification. A very good example is the PRED-GPCR server [9] where 265 signature profile HMMs have been constructed and consequently employed in the classification of GPCR sequences. Intended for predicting if a given sequence is a member of the GPCR family, it is not optimal to perform subfamily level classification. Yet, it demonstrates the use of HMMs in GPCR classification. As a consequence of using HMMs, the classification technique is very opaque and it is not straightforward to discover the key ligand interacting sites of the receptors from the profile.

There are a number of metrics that have been used in sequence analysis literature to make classification efforts more successful. A technique employed in the work of Cui et al. in [10] is to construct a feature vector for representation of the structural and physico-chemical properties of an amino acid. The amino acids in a sequence are divided into 3 categories, namely hydrophobic (CVLIMFW), neutral (GASTPHY), and polar (RKEDQN). Each of these groups is described by three descriptors, namely composition (C), transition (T) and distribution (D). These capture the amino acid composition of a sequence in 21 parameters (1 value for the composition, 1 value for the transition and 5 values for the distribution, for each category and there are 3 categories). This abstracted representation of an amino acid sequence has also been used in some very recent GPCR classification studies [11].

Similarly Atchley et al. in [12] have defined around 500 amino acid attributes which have been summarized into five continuous attributes through multivariate statistics. Such techniques which summarize the amino acids of a sequence in a number of continuous parameters are easier to integrate with many of the pre-existing classification tools or algorithms. However, such methods which summarize the entire sequence in a number of numeric metrics fail to pinpoint specific residues which are important in determining key ligand receptor interaction sites. Therefore, in order to identify the potential ligand-receptor interaction sites, these techniques were not used.

There have been numerous motif-based approaches to GPCR classification. In [13], the functions of a number of orphan receptors were predicted through multiple alignments of Class A GPCRs. In [14], [15], Chou et al. demonstrate the relationship between the amino acid composition of a GPCR sequence and its type within the amine subfamily. Another motif-based approach is to use GPCR "fingerprints" that are specific to the GPCR seven-helices structure [16], [17]. This method entails the use of well-conserved short sequence bursts that correspond to the loops, trans-membrane regions or the termini of the GPCR. The fact that each fingerprint is derived from different regions of the GPCR makes it more robust to error. The more than 270 fingerprints found in the PRINTS database allow for protein signatures to be developed for different levels of the GPCR superfamily [18]. The authors of [19] have combined the different kinds of motifs and used a swarm intelligence rule extraction algorithm to create classification rules. A more detailed description of these motif-based and other types of GPCR classifiers can be found in [20].

A recent technique, proposed in [1], entails a different approach than others to GPCR classification. A multitude of classification algorithms (10 in total) are tested at each level of the GPCR classification hierarchy and the algorithm which performs best at each level is chosen. Classification of a sequence across the GPCR hierarchy is handled by the best classification algorithm at that particular level as it progresses down the classification tree. Despite combining the strength of different classification algorithms, the downside of this work is that the classification method is very opaque. For sequence representation, 26 physico-chemical properties are selected on which they have applied Principal Components Analysis (PCA) and selected the best 5 components. Therefore, neither the sequence representation nor the classification algorithms are able to give us detailed information about which particular property of the sequence has led to the reported class prediction. This method cannot even give us a very clear perception of which physico-chemical component is most helpful because PCA combines all of them in order to produce its components.

The GRIFFIN project, which aims to predict GPCR - G protein coupling, employs an SVM-HMM hybrid which combines the efficiency of HMM with the predictive power of SVM in a SVM-HMM hybrid [21]. Most sequences are classified using HMM at the first stage which is significantly more efficient than SVM. However, when HMM fails to make a classification for the families or subfamilies for which it has been specifically trained, it passes the data on to an SVM. This SVM model (at the second stage) uses some other features and makes a classification based on them. If it fails to make a sufficiently confident guess, there is a second SVM which also looks for a parameter and makes the final decision about that sequence. A similar SVM/HMM hybrid classifier is not appropriate for the planned approach because one of the goals is to determine the key ligand-receptor binding sites with clear motifs. This classification approach cannot give clear-cut rules about why it makes certain classifications hence is eliminated as an option in this study.

The prevailing picture from these articles is that in the trade-off between transparency (i.e. the classifier's ability to report which characteristics of the input determines the classification) and accuracy, most pre-existing GPCR classification tools have shifted heavily towards accuracy. The contribution of this thesis is to propose a GPCR classifier which maintains a high degree of transparency while achieving classification accuracy that is at least as good as the preexisting classifiers. The method proposed can pinpoint possible ligand-receptor interaction sites for each subfamily of the pharmaceutically significant Class A receptors.

3 Preliminaries and Problem Definition

In section 3.2, background information on the GPCR proteins and their structural properties is given. In section 3.3, the formal definition of the GPCR classification problem is provided. In section 3.4 the various amino acid grouping schemes are introduced.

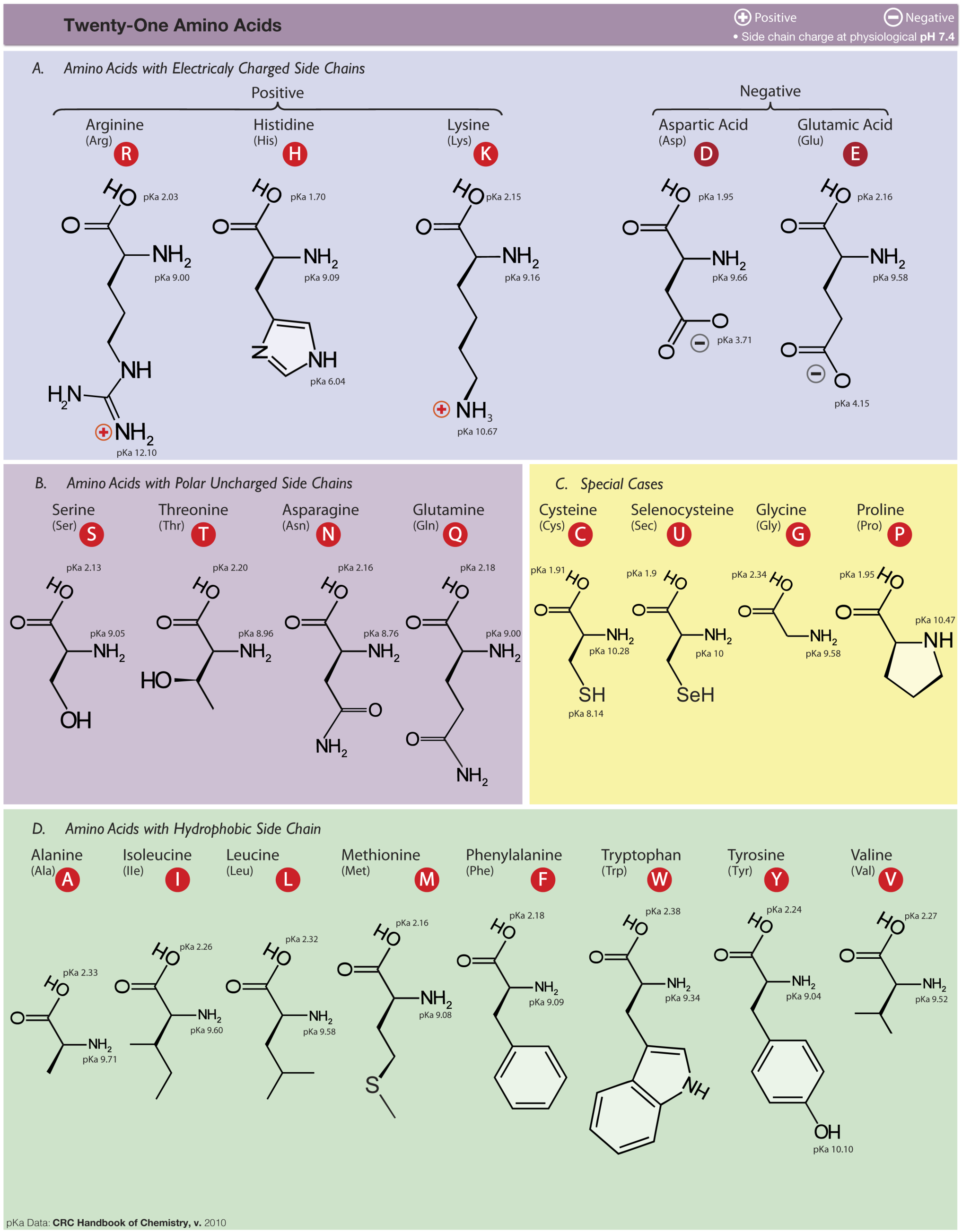
3.1 Background on Proteins

Proteins are organic polymers that are made up of amino acids connected by peptide bonds. Proteins carry out most of the functions within the body. They are made up of a chain of amino acids that fold and take different shapes. The sequence of the amino acids in a protein is mainly determined by the encoding DNA sequence. There are 20 standard amino acids with different physico-chemical properties. The amino acids and their properties are summarized in Figure 3.1. The proteins are vital to the healthy functioning of humans and most other known organisms. For humans and most other developed species, proteins are essential in almost every aspect of life from metabolism to immune responses to signal transduction (GPCR proteins perform signal transduction).

The amino acids can be clustered together depending on different properties. Depending on the type of study, different characteristics of the amino acids gain importance and therefore the properties on which the clustering is based can change. However, in general, it is possible to classify the amino acids into three broad classes: charged (negatively or positively), polar and hydrophobic as shown in Figure 3.1. During folding, the hydrophobic amino acids tend to cluster together and away from the surface of the proteins in general as most proteins function in aqueous environments. As one might expect, the oppositely charged or polarized amino acids tend to attract one another with similarly charged or polarized amino acids tend to remain apart. However protein folding is a complex procedure that is effected by a wide range of other factors as well. Protein folding is very important because the protein's structure is vital to its function. In trans-membrane proteins, as the phospholipid layer is hydrophobic, the trans-membrane regions tend to have hydrophobic helices which fit well into the membrane structure.

3.2 Background on GPCR Proteins

The largest and most diverse family of trans-membrane receptors is the G-protein-coupled receptor family. This family of receptors is activated by a diverse range of ligands or stimuli such as small peptides, amino acid derivatives, light, taste or smell [22]. The activated receptors signal the cell through G-proteins coupled to the intra-cellular region of the receptor. Due to their important role in signal transduction, more than half of the modern drugs target this particular protein superfamily [3]. The generally accepted classification for GPCRs in vertebrates is as follows: rhodopsin-like (Family A), secretin-like (Family B), glutamate-like (Family C), adhesion and Frizzled/-Taste2 [23, 24]. This hierarchy is illustrated in Figure 2. Family A is the family of highest interest from a pharmaceutical research perspective as more than 80% of all human GPCRs are in this family alone [25]. In addition the number of sequences in this family is significantly higher than the others. Therefore, the classification efforts are focused within Family A.



pKa Data: **CRC Handbook of Chemistry, v.** 2010

Dan Cojocari, Department of Medical Biophysics, **University of Toronto** 2009

Figure 1: The table of amino acids found in eukaryotes, clustered with respect to their side chain charge at physiological pH 7.4, copied from [2].

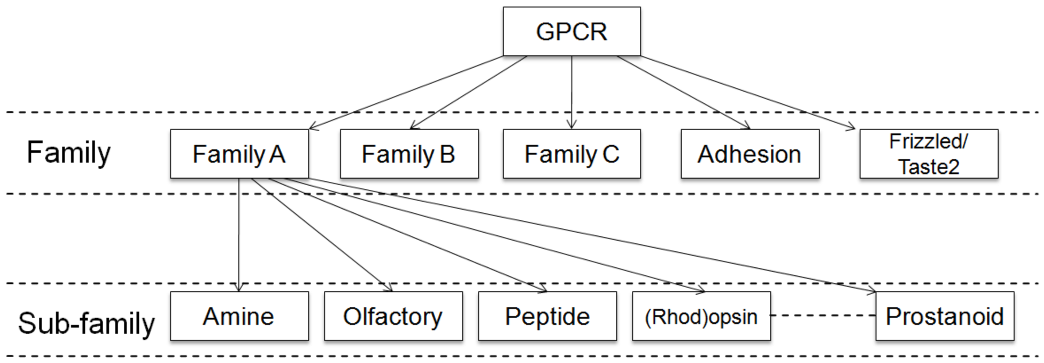


Figure 2: The GPCR classification hierarchy

Despite the significant volume of pharmaceutical research on GPCRs, the three-dimensional structures have been very hard to discover. Currently there are only four known GPCR structures in their inactive states [23]. The identification of orthosteric ligands has been similarly difficult: despite having identified more than 1000 genes encoding GPCRs, only few highly selective synthetic ligands for these GPCRs can be designed [26]. One of the reasons that identifying orthosteric full agonists has been so difficult is that G-protein activation requires various interactions at key sites between the receptor and the hormone [23]. Further complicating is that the orthosteric binding sites across members of a single GPCR subfamily are often highly conserved making specificity a major problem [26].

One of the key challenges in GPCR research is identifying these key interaction sites governing receptor agonism and conserved over the sequences in the same subfamily. These sites would be highly beneficial to drug design efforts. Another important challenge is the classification of orphan GPCR sequences. A sequence is called an orphan GPCR if it has high similarity to known and annotated GPCR sequences but nothing is known about its structure or the activating ligand. As the gap between the number of identified sequences and the number of annotated sequences grows so does the number of orphan GPCRs. Therefore, there is a strong need for successful classification of GPCR sequences especially those in the family most relevant to human drug design: Family A. This thesis is focused on classifications between the subfamilies of Family A.

An important property of the GPCRs is that certain amino-acid residues are well conserved across the family [13]. This property has been exploited in multiple studies to synthesize new GPCRs [27, 28]. The well conserved amino-acid residue property has been exploited in this study while defining the motifs.

It is also worth noting that all GPCRs share a particular structural outline. This structure, common to all GPCR sequences, is an extra-cellular amino terminus, an intra-cellular carboxyl terminus and 7 trans-membrane helices separated by intra-cellular and extra-cellular loops [23] as seen in Figure 3.

A major source of GPCR sequences is the GPCRDB [29]. The objective of the GPCRDB effort is to centrally collect and distribute all known GPCR sequences and their annotated functions. The GPCRDB contains thousands of annotated GPCR sequences and its content is easily accessible via either an interactive web-interface or easy-to-use web services. The intuition verification dataset was collected from the GPCRDB as described in Section 5.1. The performance comparison experiments are based on datasets used for training other classification servers.

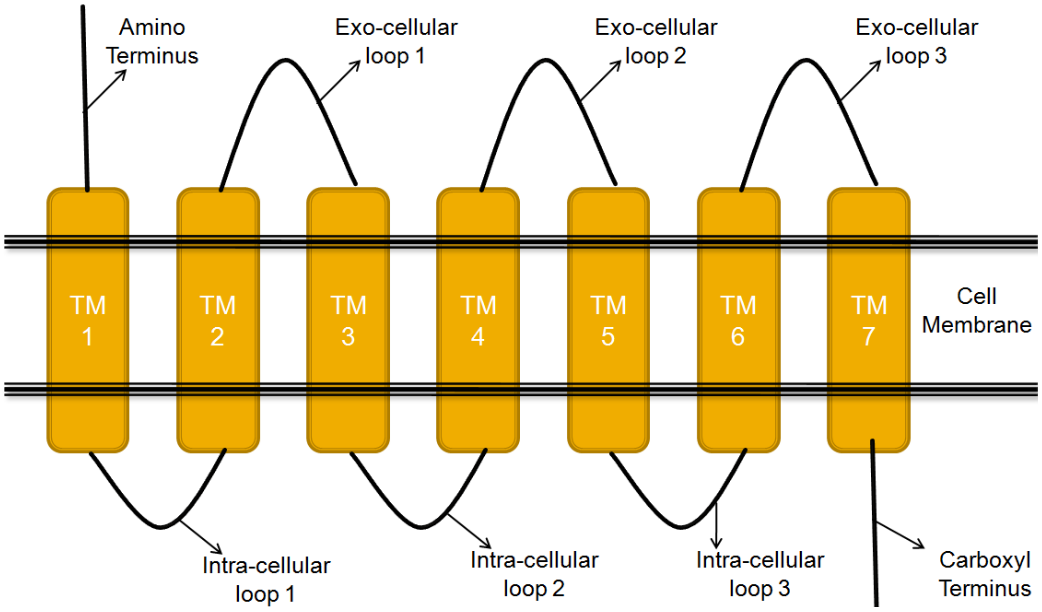


Figure 3: Representative snake-diagram of a GPCR

3.3 Classification Problem

To define the GPCR classification problem, first a formal definition of the GPCR sequence dataset needs to be given.

Definition 1 ***GPCR Sequence Dataset*** is a set of tuples (σ, χ), where

- *σ is the sequence that encodes a protein from the GPCR Family A.*
- *χ denotes the subfamily of the protein encoded by σ.*

Classification takes a training dataset whose class-membership information is utilized to extract rules for classification. This algorithm takes a testing set of sequences alone and produces the predictions for their families. The formal definition of the structure of the classification problem is defined below:

Definition 2 ***GPCR Classification Problem*** is to build a classifier C by training on the GPCR sequence dataset D which predicts the χ values of the elements of the testing dataset T.

The presence/absence of the discovered motifs are the attributes of each sequence. The classification function aims to capture the relationship between the motifs in an effort to identify the correct subfamily to which a given sequence belongs. Classifiers identify the characteristics of the data by learning the trends in the data using statistical methods. This is achieved by studying the attributes of each member of a class (in this case, subfamily) and identifying those that best distinguish one from another.

The inherent difficulty of the problem at hand is that, the attributes to be used in classification need to be discovered before being able to employ any classification algorithm. The raw data is in the form of a sequence of amino acids that constitute a GPCR protein when synthesized. Therefore an attribute/feature selection step through data mining techniques is needed. The objective of the feature selection technique is to select the attributes that are most relevant to the classification problem at hand. A novel motif evaluation metric called Motif Specificity Measure, and a motif extraction algorithm called Distinguishing Power Evaluation which uses this metric are developed.

The agonism of a synthetic ligand (drug) may not be simply associated with occupying the binding site but instead it may be determined by whether it can form the complex interactions of the endogenous ligand [23]. It is also known that the key ligand interaction sites of the receptors in a given subfamily should be well-preserved. This is pointed out by empirical data which supports that it is very hard to achieve specificity within a subfamily - i.e. what binds to one member of a subfamily often binds to all [26]. Therefore, identifying sites of ligand-receptor interaction would be important in helping drug design.

Definition 3 ***Interaction Site Identification Problem*** is to identify the amino-acid residues preserved across the sequences in the same subfamily which constitute the key ligand-receptor interaction sites.

To identify the different regions of a GPCR, it is essential to identify the trans-membrane helices. TMHMM is a widely recognized computational trans-membrane region prediction tool [30], [20]. Since the trans-membrane helices are buried in the lipid membrane, they are mostly made up of hydrophobic amino acids. These regions can be captured by hidden Markov models since their transition and emission rates show a significant difference for the helical regions of GPCR proteins. TMHMM does exactly this: it uses a hidden Markov model (HMM) to predict the position of the trans-membrane helices. When the trans-membrane helices, we have information about the extra-cellular and intra-cellular loops of a given protein sequence as well. The current version of TMHMM is 2.0 and it can be accessed at http://www.cbs.dtu.dk/ services/TMHMM/.

3.4 Amino Acid Grouping Schemes

A common practice in sequence-based studies is to reduce the 20-letter alphabet to a smaller number by grouping the amino acids together. The most significant benefit of reducing the amino acid alphabet is that it creates a smaller set of possible motifs. This reduces the search space of all motifs, making classification more robust to random changes in the DNA. Certain amino acids with similar physico-chemical properties could replace one another during these random changes without disturbing neither the protein structure nor function such as, Isoleucine, Leucine, Valine and Alanine. By generalizing similar amino acids into a single group and representing all of them with a single letter in the reduced alphabet, more robust motifs that are less prone to error in the face of evolutionary DNA changes can be identified.

An important problem here is to define which amino acids can be considered similar. There are a number of basic physico-chemical properties such as hydrophobicity, charge, mass etc which can be used as a basis of grouping but any such attempt needs to prioritize over some others to perform a successful grouping. It should also reduce the number of clusters to a small number to be worth using any reduction scheme at all. Given these restrictions, a reduction table to optimize the capability to capture GPCR binding properties had previously been designed and used in [8]. There is previous work by Davies et al. [31] which focuses exclusively on optimizing these amino acid groupings. The grouping schemes taken from this paper were those that were found by the highest cross-validation fold for both the seeded and random initialization techniques. Finally, a small adjustment to the Davies seeded reduction scheme was made to create Davies seeded 2, resulting in four different amino acid reduction schemes as shown in Table 1. In this table, each amino acid is represented by its single-letter code. Sezerman's grouping gave the best results and was used in the rest of the study.

In order to see the effects of grouping, experiments without any grouping

Grouping Scheme	A	B	C	D	E	F	G	H	I	J	K
Davies Random	SG	DVIA	RQN	KP	WHY	C	LE	MF	T		
Davies Seeded 1	SGE	DP	RWN	KQ	HLVIMFY	C	AT				
Davies Seeded 2	SGE	DP	RWN	KQH	LVIMFY	C	AT				
Sezerman	IVLM	RKH	DE	QN	ST	A	G	W	C	YF	P

Table 1: The amino acid grouping alternatives tested.

were carried out as well. Unless the grouping schemes provide a significant boost to the accuracy of the classifications - hence the confidence of the conclusions - no grouping techniques are superior, because using a grouping scheme blunts the quality with which the interaction sites are identified. The information content of non-reduced motifs is higher; therefore, they are preferable to any grouping scheme in case the respective distinguishing abilities are comparably powerful.

Sezerman grouping gave the best results among these alternatives; therefore, all results reported will be according to Sezerman's grouping.

4 Method

The method proposed in this thesis to solve the classification problem described above can be summarized as follows:

1. Motif distillation by Motif Specificity Measure (Motif definition is in 4.1 and MSM definition is in 4.2)

2. Distinguishing Power Evaluation of distilled motifs

3. Decision Tree induction from selected motifs

4. Identification of key ligand interaction sites through rule extraction from decision tree.

5. Classification of subfamilies using "key ligand interaction site motif" presence.

The classification rules are simply rules dictating the presence or absence of some motifs. The design of the motifs allows us to predict ligand interaction sites from sequence information alone. Throughout this section, the term class will be used to denote the subfamily to which a sequence belongs for the sake of simplicity. As the classification problem is single-level, this should not create any ambiguity.

4.1 Motif Definition

Sequence information in its raw form – without feature extraction – cannot be used to perform any classification. Machine learning algorithms are more effective when the input data have few but distinguishing attributes. Therefore, extracting distinguishing motifs from the sequence information would positively effect the accuracy of supervised learning methods in general. The motifs are also required to clearly represent some location-specific properties of the sequences because the objective of this study is two-fold: to determine key interaction sites as well as perform classification. This requirement has led us to depart from the other motif definitions in literature such as [10, 1] and define a novel motif.

The intuition was that within a subfamily, certain amino-acid triplets at specific positions of the same exo-cellular region would be preserved over the different sequences in the subfamily. This intuition is illustrated in Figure 4: the ligand that binds to the receptor interacts very strongly with a number of key sites (highlighted in blue), which is captured by the motif definition. It can be speculated that these amino-acids might be fundamental to the binding process because otherwise they would not have been conserved. As there is not a sufficiently large number of GPCR structures to determine location in a spatial sense, the use of the word location from here on refers to a sequential location. Sequential location means the location of the amino acids within the entire sequence; a linear sense of positioning where the start is the first amino acid of the sequence and the end is the last amino acid in the sequence. With location defined as such, the conserved sites should be excellent motifs for classification if the intuition holds. If conserved sites point to key interaction sites in the binding process the motifs of one subfamily should not occur in another subfamily – otherwise the same ligands would bind to receptors of both subfamilies and they would be classified in the same subfamily. This intuition is experimentally verified in section 5.1. The motifs are designed with this intuition.

Definition 4 ***Motif Definition*** The motif is defined as $m(\tau, r, p)$ where

- τ *is a triplet of residues from the preferred amino acid alphabet.*
- r *is the exo-cellular region of occurrence, where it is one of the following: n-terminus, exo-loop 1, exo-loop 2 or exo-loop 3.*
- p *is the position of the first residue of the triplet relative to the length of the amino acid sequence of region* r*.*

In a previous work, it is expressed that features of length three are the most informative for classification of GPCR sequences [32]. The study uses an SVM-based classifier for performing GPCR Class A subfamily-level classifications. Therefore, the reported fact that features of length three are the most informative is valid for this study as well as other Class A subfamily classification studies.

To determine the trans-membrane regions, the TMHMM trans-membrane helices prediction tool was used[30]. The trans-membrane regions can be predicted with high accuracy due to the very significant difference in hydrophobicity with the extra-membrane regions. The TMHMM tool was picked over other alternatives because a comparative study has found it to be the best among a suite of tools that perform the same prediction [33]. Once the trans-membrane regions are identified, it is trivial to identify the exo-cellular regions. The term region here refers to one of the four exo-cellular components which are common to every member of the GPCR family. These exo-cellular

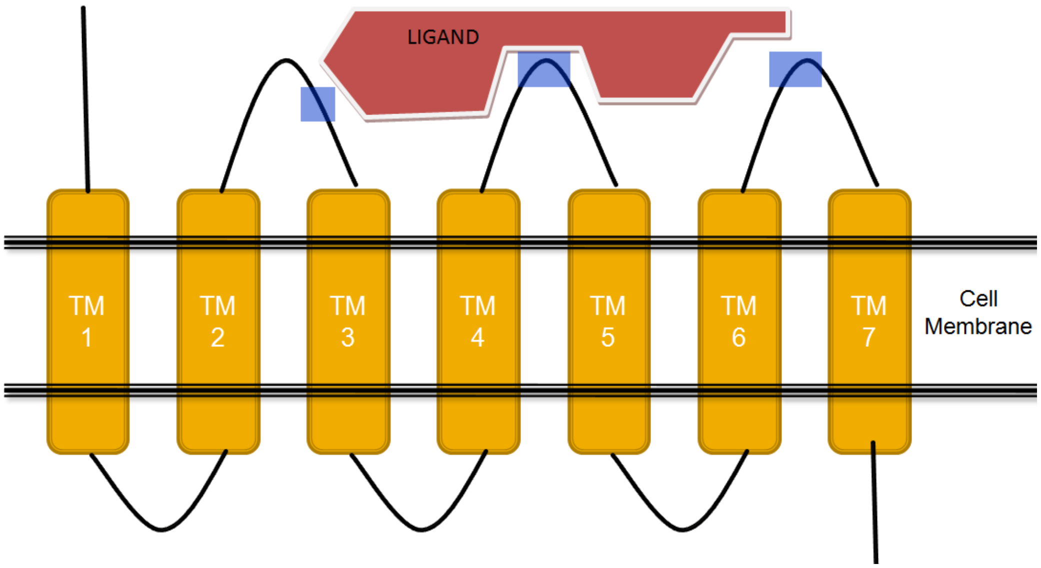


Figure 4: Illustration explaining the inspiration for motifs.

components are n-terminus, exo-loop 1, exo-loop 2 and exo-loop 3 as can be seen in Figure 3. The regions are 0-indexed such that the n-terminus region is indexed 0, the exo-loop 1 region is indexed 1 etc.

For a motif $m(\tau, r, p)$, the position within the region is defined to be the sequential position of the first letter of the triplet within the loop, normalized by the length of the loop. This allows us to define the notion of position independent of the length of the region. For example a triplet appearing in the middle of a region of size 10 and a triplet occurring in the middle of a region of length 50 have the same relative position although one of them starts at index 5 and the other starts at 25. This maps the position of a triplet from a number with an indefinite range (which varies as the number of residues in the loop changes) to a number between 0 and 9. The position was limited to integers between 0 and 9 because empirical study revealed that the average region length was 26.5 for the GPCRpred dataset. As the residues are evaluated in consecutive strips of length three, the number of disjoint triplets is around 10. Exact calculation of a position is given in Definition 5 which is illustrated by Example 1.

Definition 5 ***Position Calculation*** For position p in region r, the triplets that occur in that position start with index $\left\lfloor p \times \frac{|r|-1}{10} \right\rfloor$ where $|r|$ denotes the sequence length of region r and the residue indices start from 0. The beginning residue of the first segment is the first residue (index 0). The end of a position segment is the first residue of the next segment or the end of the region if this is the last segment. The residues that occur in the such defined region constitute the first residues of the triplets in that position where the rest of the triplet is simply the two consecutive residues.

Example 1 ***Calculating triplet positions*** *Assume that a region consists of the following 19 residues: "ARNDCEQGHILKMFPSTWY". The triplets at position 3 can be calculated by filling in the necessary values to the formula specified in definition 5.* $\left\lfloor 3 \times \frac{19-1}{10} \right\rfloor = 5$*. The beginning of the next position (i.e. position 4 is calculated similarly:* $\left\lfloor 4 \times \frac{19-1}{10} \right\rfloor = 7$*. The triplets that are in position in 3 start with indices in the range* $[5, 7)$ *– in other words the triplets that start with the indices 5 and 6 fall in position 3. Therefore, the triplets that occur at position 3 of this region are EQD and QGH. Assume that the given region is the n-terminus region of a sequence, then it can be said that the motif* $m(QGH, 0, 3)$ *occurs in this sequence. Table 2 shows the starting index of each position and the triplets belonging to each position segment for a region with the following sequence of length 19: "ARNDCE-*

Table 2: The triplets in each position of the region "ARNDCEQGHILKMFPSTWY"

Position	Triplets in this position start with index	Occurring Triplets
0	0	ARN
1	1,2	RND,NDC
2	3,4	DCE,CEQ
3	5,6	EQG,QGH
4	7,8	GHI,HIL
5	9	ILK
6	10,11	LKM,KMF
7	12,13	MFP.FPS
8	14,15	PST,STW
9	16	TWY

QGHILKMFPSTWY".

4.2 Motif Specificity Measure

The total number of motifs is on the order of hundred thousands; however, most of them occur very infrequently. The ideal motif would be one that occurs in all the sequences that belongs to a particular subfamily but never in a sequence from another subfamily. To evaluate how close a motif is to this ideal, the metric should give a high value for motifs that occur frequently in one subfamily but are very uncommon in other subfamilies. This way, motifs that are specific to a particular subfamily would be rewarded whereas motifs which occur either in few sequences or in multiple subfamilies would be penalized.

Metrics with similar properties are used in the field of text mining. The numerous words which occur in every text cannot be used for efficient document retrieval instead the most specific words in a query need to be selected. The Term Frequency Inverse Document Frequency (TFIDF) [34] weight is a metric that selects words with high occurrences in a low number of documents. The weight increases as the occurrences of a word in a document increases; however, it is inversely proportional to the number of overall documents in which the word occurs. This allows the weight to be high for those words that are specific which is highly similar to the sought-after characteristic of the Motif Specificity Measure. Therefore, the TFIDF weights were the starting point in defining the Motif Specificity Measure.

The Motif Specificity Measure of a motif is composed of two components, the first of which is directly proportional to the motif's presence in the target subfamily.

Definition 6 ***Presence in Family*** Presence of motif i in family f, $PF(i, f)$ is given by

$$PF_{i,f} = \frac{n_{i,f}}{\sum\limits_{k \in M} n_{k,f}} \tag{4.1}$$

where

- *$n_{i,f}$ is the number of occurrences of motif i in unique sequences in subfamily f,*
- *M is the set of all motifs,*
- *$\sum\limits_{k \in M} n_{k,f}$ denotes the total number of occurrence of all motifs in subfamily f.*

The second component is the Family Specificity of a motif which is inversely proportional to the number of different in which that particular motif occurs. Here, deciding the occurrence of a motif in a subfamily is not trivial. Occurrence of a motif in a single sequence out of hundreds of sequences in a subfamily is hardly the same as a motif to be observed in more than half of the sequences of a subfamily. Occurrence of a motif in a single sequence in an entire subfamily can be due to numerous reasons such as wrong sequence annotation, evolutionary connections etc. Therefore, a motif is said to occur in a subfamily only if its occurrence rate in the subfamily is higher than a certain percentage threshold, called the *Presence Threshold.*

Definition 7 ***Motif Occurrence Rate in a Family*** The occurrence rate of motif $i(\tau, r, p)$ in subfamily f, $MORF_{i,f}$ is given by

$$MORF_{i,f} = \frac{\sum\limits_{s \in f} |Occurs(i,s)|}{|f|} \tag{4.2}$$

where

- *$Occurs(i,s)$ evaluates to 1 if motif i occurs in sequence s, otherwise 0,*
- *$|f|$ is the number of sequences in subfamily f*

Given the motif occurrence rate in a subfamily, the Family Specificity can be defined as follows:

Definition 8 ***Family Specificity*** The Family Specificity of motif i, FS_i is given by

$$FS_i = \log \frac{|F|}{\sum\limits_{f \in F} |\{f : MORF_{i,f} > PT\}|} \tag{4.3}$$

where

- *F is the set of all subfamilies,*
- *$MORF$ is the Motif Occurrence Rate in Family function defined above,*
- *PT is the Presence Threshold.*

The denominator of FS simply gives the number of subfamilies for which the occurrence rate of a particular motif is above the Presence Threshold. The reason the Presence Threshold is introduced, is to be able to cope with subfamilies of very different sizes. In this case, with the standard method of calculating IDF score, the total number of sequences outside the target subfamily needs to be divided with the total number of sequences outside the target subfamily in which the motif has been seen. This would have treated presence in every sequence equally – regardless of its subfamily. More often than not, the number of sequences in different subfamilies differ greatly – sometimes even by one order of magnitude. Therefore, if a motif showed significant occurrence in only one very large subfamily, its FS score would have been equal to that of a motif which shows significant occurrences in many subfamilies with smaller number of sequences. However, the specificity of the two motifs are hardly the same: the former occurs frequently in only one subfamily outside its target subfamily whereas the latter occurs in many different subfamilies. To cope with subfamilies of very different sizes the number of subfamilies in which the motif occurs frequently, where "frequent" is determined by the *Presence Threshold*, are counted. The value of Presence Threshold should not be too high so that motifs with frequent occurrences in a subfamily should be noted. However, it should also be high enough to prevent minor motifs from appearing significant. The best trade-off was assessed to be at the 20% level and this value was used in the computations.

The Presence in Family and the Family Specificity of a motif enable us to capture two key properties in assessing the specificity of a motif to a subfamily. The Motif Specificity Measure which determines the specificity of a motif to a particular subfamily is then defined as follows:

Definition 9 ***Motif Specificity Measure*** The Motif Specificity Measure of motif i for subfamily f, $MSM(i, f)$ is given by

$$MSM(i, f) = PF_{i,f} \times FS_i \tag{4.4}$$

where

- $PF_{i,f}$ *denotes motif* i*'s Presence in Family* f*,*
- FS_i *denotes the Family Specificity of motif* i*.*

The Motif Specificity Measure of a motif for a particular subfamily is positively correlated with the number of occurrences of a motif in that subfamily but inversely correlated with the number of other subfamilies in which the motif occurs frequently.

4.3 Distinguishing Power Evaluation

In the Distinguishing Power Evaluation (DPE) step, the training data is used to determine the best motifs for classification. The central idea is to repeatedly build decision trees from randomly partitioned test and training data and look for those motifs that occur very frequently in each of these decision trees. The aim of the DPE algorithm is not to produce a classifier but rather evaluation of the motifs via a thorough analysis of the data. The flowchart of GPCRBind is shown in Figure 5.

During the DPE step, the Distinguishing Power (DP) score of each motif, which is simply the sum of the accuracies of the decision tree in which that motif occurs, is calculated. If a motif occurs in many decision trees which performed high accuracy classification, then using that motif as an attribute yields a significant information gain. This is due to the characteristic of the Iterative Dichotomiser 3 (ID3) decision tree induction algorithm [35] which splits the data with respect to the information gain of the attributes. The ID3 algorithm uses an attribute at a decision tree node only if this attribute yields the highest information gain at that node of the tree.

The first part of the DPE is to filter the number of candidate motifs from hundreds of thousands to hundreds. Initially every triplet, region and position combination is a candidate motif. However, most of these motifs occur extremely infrequently whereas some of the rest occur in most GPCR sequences as they are characteristic to the subfamily. Neither of these types of motifs would contribute much information to help solve the classification problem. Therefore, the motifs with the highest subfamily specificity are picked using the MSM which has been described in section 4.2. Algorithm 4.3 details the procedure for elimination of motifs using MSM, shortly ElimSM.

To understand Algorithm 4.3, it must be underscored that a motif's MSM can only be evaluated with respect to a subfamily, since the MSM score gives clues about how useful each motif will be for the classification of that particular subfamily. For each subfamily, N motifs with highest MSN scores

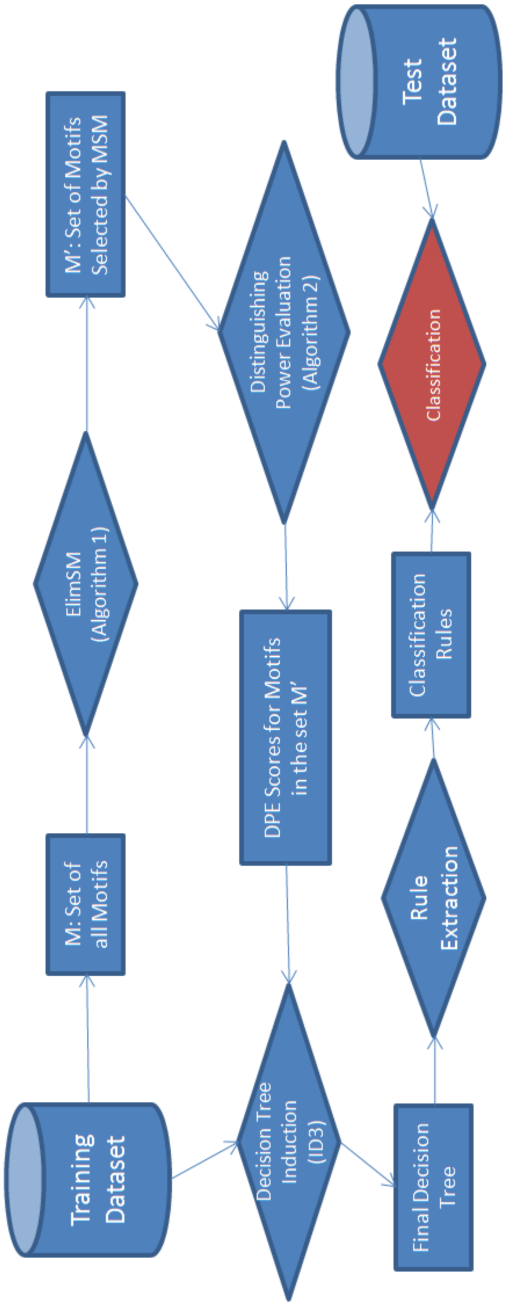

Training Dataset
M: Set of all Motifs
ElimSM (Algorithm 1)
M': Set of Motifs Selected by MSM
Distinguishing Power Evaluation (Algorithm 2)
DPE Scores for Motifs in the set M'
Decision Tree Induction (ID3)
Final Decision Tree
Rule Extraction
Classification Rules
Classification
Test Dataset

Figure 5: The flowchart of GPCRBind.

Algorithm 1 Calculating Motif Specificity Measure (ElimSM)

Input: Set of motifs M, set of subfamilies F, cutoff value N.

Output: Set consisting of N motifs with the highest MSM value for each subfamily

```
$BestM \leftarrow \{\}$
for all $f \in F$ do
    $BestM_f \leftarrow \{\}$
    $Scores_f \leftarrow MSM(M, f)$
    for all $m \in M$ do
        //If $m$ is among the top scoring motifs for this subfamily, add it to the corresponding set of best motifs.
        if $MSM\,(m, f)$ in $MaxN(Scores_f)$ then
            $BestM_f \leftarrow BestM_f \cup m$
        end if
    end for
    $BestM \leftarrow BestM \cup BestM_f$
end for
return $BestM$
```

where;

- $MSM(M, f) = \{MSM(m, f) : m \in M\}$
- $MaxN$ takes as input a set with a score assigned to each element and returns the N highest scoring elements of this input set.

have been selected. Since N is a natural number, the value of N is determined automatically in a hill-climbing manner by sampling the alternative cutoff values on a training set and then selecting the value that yields the highest accuracy. The value of N is calculated dynamically for every dataset to make sure that the algorithm can adapt to datasets with different characteristics.

In order to maximize the strength of decision trees a sufficiently good set of attributes of each data object, which distinguishes between the various subfamilies, needs to be given. In this study, the data objects are the sequences and their attributes are defined to be the presence of the motifs selected through the MSM elimination step. Each sequence has as many attributes as the number of selected motifs which is equal to number of subfamilies multiplied by the number of motifs per subfamily (the value N in algorithm 4.3). Each attribute is a binary attribute denoting the presence/absence of the corresponding motif. If the corresponding motif of an attribute occurs in a sequence then the value of that attribute is 1 for that sequence, otherwise it is 0. The dataset of GPCR sequences can thus be converted into a classification-ready dataset as defined in 10.

Definition 10 ***Classification Dataset*** The classification dataset C is created from a GPCR sequence dataset D and a set of motifs M such that;

- $\forall s \in D, \exists s' \in C$,
- $\forall s' \in C$ *has as many attributes as* $|M|$,
- $s'_i = 1$ *if* $m_i \in M$ *occurs in sequence* s,
- $s'_i = 0$ *if* $m_i \in M$ *does not occur in sequence* s.

The DPE algorithm (Algorithm 4.3) is, in its essence, a reiteration of decision tree building. Initially the DPE score of all motifs is 0. As the various decision trees are built and tested from random partitions of the training data, the resulting accuracy of each tree is added to the DPE score of every motif on that tree. If there are multiple occurrences of a motif in a single tree, the DPE score is incremented only once. This ensures that the motifs with high DP scores are those motifs that occur in a high number of trees and in high accuracy trees.

The varying factor over the iterations of the DPE algorithm is the data partitions. At each iteration, the input data of the algorithm is randomly divided into three partitions. One of these partitions is dedicated as the test set and the remaining partitions are merged to form a training set. The motif elimination by MSM step is done using the training set only and the best motifs which explain the training set are derived. The training and test sets are converted into classification datasets where the attributes are the motifs selected in the previous step. The test set can be converted to a classification dataset format as well because the conversion only requires the sequence, not the class information. The next step is to train a decision tree on the classification-format training set using the ID3 algorithm and classify the test set using this decision tree. The accuracy of the tree on the test set is added to the DPE score of every motif used in the decision tree. The reported results have been achieved by using 20 runs.

Algorithm 2 Distinguishing Power Evaluation

Input: Sequence Dataset D

Output: Motifs and corresponding DPE scores

```
$\forall m \in M, DP_m \leftarrow 0$
**for** $run = 1 : TotalRuns$ **do**
    $F \leftarrow$ Retrieve subfamilies from $D$
    $P = \{P_1, P_2, P_3\} \leftarrow RandomPartition\,(D)$
    **for all** $P_i \in P$ **do**
        $TestSet \leftarrow P_i$
        $TrainSet \leftarrow P/P_i$
        $M \leftarrow FindAllMotifs\,(TrainSet)$
        $BestM \leftarrow elimSM(M, F, N)$
        $C\_train \leftarrow ClassDataset(BestM, TrainSet)$
        $C\_test \leftarrow ClassDataset(BestM, TestSet)$
        $decisionTree \leftarrow ID3(C\_train)$
        $accuracy \leftarrow decisionTree.Test(C\_test)$
        **for all** $m \in BestM$ used in $decisionTree$ **do**
            $DP_m \leftarrow DP_m + accuracy$
        **end for**
    **end for**
**end for**
```

4.4 Discovery of Key Ligand Interaction Sites

As one of the objectives is to identify the key ligand-protein interaction sites, the classification method being used should produce clear, direct yet powerful rules for each class. The decision trees are tools that could be used for extracting such rules and it was decided that the Iterative Dichotomiser 3 (ID3) algorithm proposed by Quinlan [35] is the best alternative. ID3 is a simple yet powerful algorithm; its output is a decision tree which can be parsed for the important rules which in turn yield high accuracy results. The rule generation algorithm also serves to prune the decision tree, counteracting over-fitting which can be considered one of the major downsides of ID3-based decision tree induction.

The DPE score characterizes the distinguishing power of a motif, as its name implies. Therefore, motifs with low distinguishing power are eliminated before extracting classification rules. The maximum possible DPE score of a motif is the score that a motif would have if it occurred in all the decision trees generated in the DPE algorithm and if all of these decision trees had 100% accuracy. Motifs with DPE scores below a threshold percentage of this maximum DPE score are eliminated. For example, a 10% threshold implies that motifs with less than 10% of the maximum possible DPE score are eliminated. This threshold is called the DPE motif selection threshold and its effect on runtime and accuracy is explained in Section 5.4.

The reason that the motifs who fall below the specified threshold are eliminated is that these motifs have either occurred in few trees or they have occurred in many trees with very low accuracies. Both rarely selected motifs and motifs that have occurred in unsuccessful trees are poorly performing motifs; therefore, they are eliminated.

The motifs that pass the DPE motif selection threshold are picked as the attributes of each sequence for the induction of the final decision tree. The whole training set is used to build the final decision tree. The selected motifs with the highest DPE scores are used to create the final decision trees using the entire body of training data available. One decision tree is produced which, given a GPCR sequence, predicts the subfamily to which it belongs.

The final decision tree is then used to extract rules as described by Quinlan in [36]. First, each path from the root of the decision tree to the decision nodes at the leaves are traced. The path is a sequence of nodes where each of these nodes represents a different attribute - therefore, by definition of the attributes, the existence of a motif. All the nodes visited until a leaf node form a set of conditions upon which a particular classification is made. The conjunction of the conditions that need to be met to reach a particular classification decision constitutes a classification rule. The conditions of these classification rules can be simplified by dropping the useless conditions. The least relevant condition to the classification is found using Fisher's Exact Test [36] at 99% confidence level. This process is repeated until there are no conditions left or there are no conditions which can be rejected at this significance level. Each of these rules are assigned a confidence factor (CF) which measures how many members that satisfy the conditions of the rule actually belong to the class proposed by the rule in the training set.

To be able to use Fisher's exact test, an appropriate alpha value had to be selected. High alpha values would involve too many motifs; therefore, over-fitting the training set to possibly reduce performance on a blind dataset.

Too many motifs would also make it more difficult to separate very significant interaction sites from those not as common. Given the above considerations and the sensitivity of biological data, the tests were performed at the 1% significance level.

After the conditions have been simplified, the rule set is evaluated as a whole in terms of the degree of success in the absence of each rule. If the rule set performs better or equally well when one of the rules is removed, the rule whose absence increases the performance the most gets eliminated, and the analysis is repeated.

Classification of a sequence is decided by the rule for which the sequence matches all the conditions. If there are more than one of such rules, then the rule with the highest confidence factor is picked. If the confidence factors are equal as well, the rule with more conditions is preferred on the grounds that it is more specific.

The classifier is the entire rule set determined as described above. Each rule is composed of conditions which dictate the presence/absence of one or more motifs. Here it should be noted that compliance with the "motif presence condition" requires that a particular motif occurs in a sequence. Similarly "motif absence condition" requires that the motif does not occur in a sequence.

A rule composed entirely of motif absence conditions would not be of much use or would not contribute a lot of information to the drug designers. However, a rule with all of its conditions being absence motifs fails to pass the Fisher's Exact Test statistical threshold simply because they appear in too many different subfamilies and are hardly unique to one class. Therefore, rules made entirely of motif absence conditions are dropped by the algorithm. As a result, the design of the proposed technique is such that it ensures there is at least one motif presence condition in any derived rule.

The classifier proposed here is called GPCRBind. The performance of the GPCRBind classifier is reported in Section 5.

5 Experimental Results

The proposed techniques were implemented in Python 2.5 and tested their performance on real datasets and compared its performance to state-of-the-art GPCR classifiers. The experiments were performed in a server with 6 Intel Xeon 2.4Ghz CPUs, 32 Gb of memory and CentOS 5.4 operating system.

The first set of experiments were conducted to verify the motif definition as presented in Section 5.1. This verification step showed that the motif definition can accurately identify GPCR subfamily-specific features. In Section 5.2, the classification performance of the proposed method is evaluated. The performance evaluation has been conducted in two steps: performance comparison between an existing classification server, GPCRpred, and the method is given in Section 5.2.1; the performance evaluation on an independent dataset and its comparison to the GPCRTree and PRED-GPCR methods is given in Section 5.2.2. The accuracy-runtime trade-off is explained in detail in Section 5.4. The discovered interaction sites are presented in Section 5.5.

5.1 Verification of the Motif Definition

The intuition while defining the motifs was that there would be certain conserved sequences in the extracellular regions of the receptors. If the intuition holds, the technique must be able to identify motifs with very high occurrence rates at certain positions for each subfamily. If there are such conserved motif occurrence patterns, then this means that these motifs can be utilized for classification. To verify this intuition experiments were made on a dataset consisting of five subfamilies of the Class A GPCRs: Amine (561 sequences), Peptide (1291 sequences), Rhodopsin (643 sequences), Prostanoid (83 sequences) and Olfactory (2311 sequences) from the GPCRDB database.

A statistical analysis of occurrence for every possible motif was performed and the occurrence positions were plotted on a histogram. The x-axis of the histogram represents the position of occurrence of the triplet within the region. The y-axis represents the number of occurrences. If the intuition is correct, there should be at least some amino-acid triplets which cluster around a few positions with extremely high occurrence rates. The analysis did indeed show that there were such occurrences and this has – to some extent – verified the intuition. You can see the histograms of such nature with the Sezerman amino acid reduction scheme in Figures 6 to 10. What is even more significant is that these motifs are those with the highest Motif Specificity Measure scores. Therefore, these data-derived results verify the intuition behind the motif definition and demonstrate the effectiveness of MSM.

5.2 Classification Results for Subfamilies of Class A

The performance of GPCRBind was compared against the literature on both an independent training set and against a GPCR classification server. The independent dataset testing is essential to show its performance when it performs on data that it has not previously encountered. Most often, when GPCRBind is used to perform classification of GPCR sequences, they will be novel sequences and it is imperative to test the performance on such data beforehand.

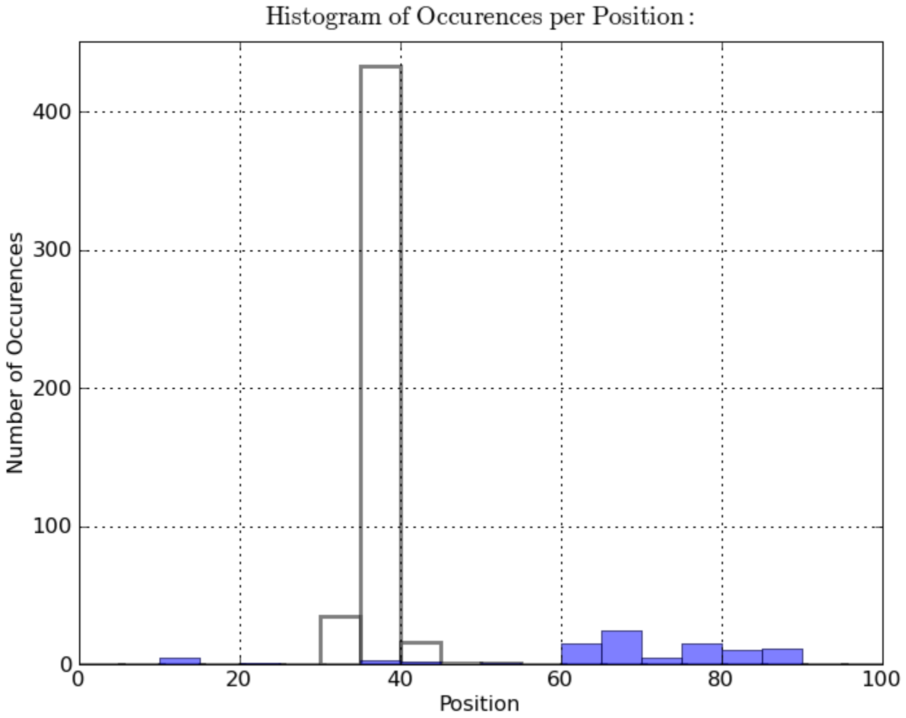


Figure 6: The occurrence frequency of triplet EIG at exo-loop 2 in rhodopsin subfamily (represented by white bars) and the other subfamilies (represented by blue).

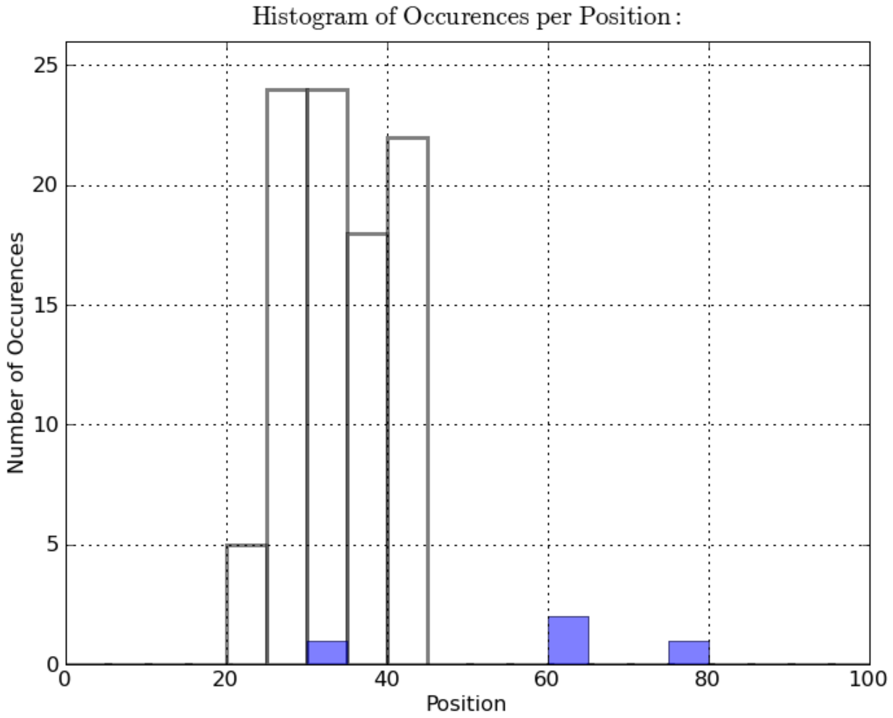


Figure 7: The occurrence frequency of triplet EHI at exo-loop 2 in prostanoid subfamily (represented by white bars) and the other subfamilies (represented by blue).

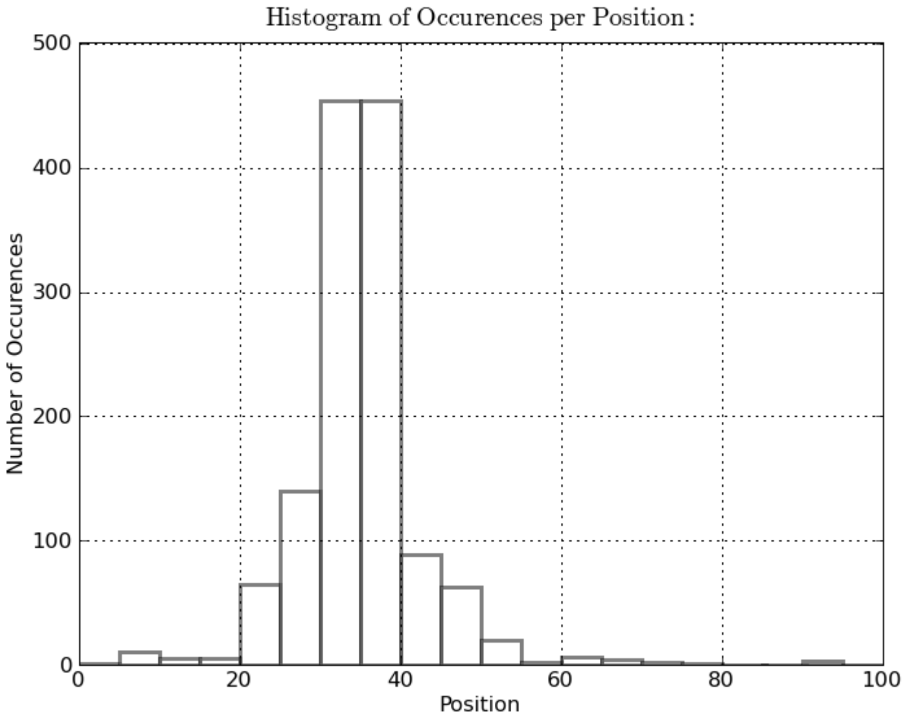


Figure 8: The occurrence frequency of triplet JJI at exo-loop 2 in olfactory subfamily (represented by white bars). The other subfamilies are so insignificant that they are not visible in the histogram.

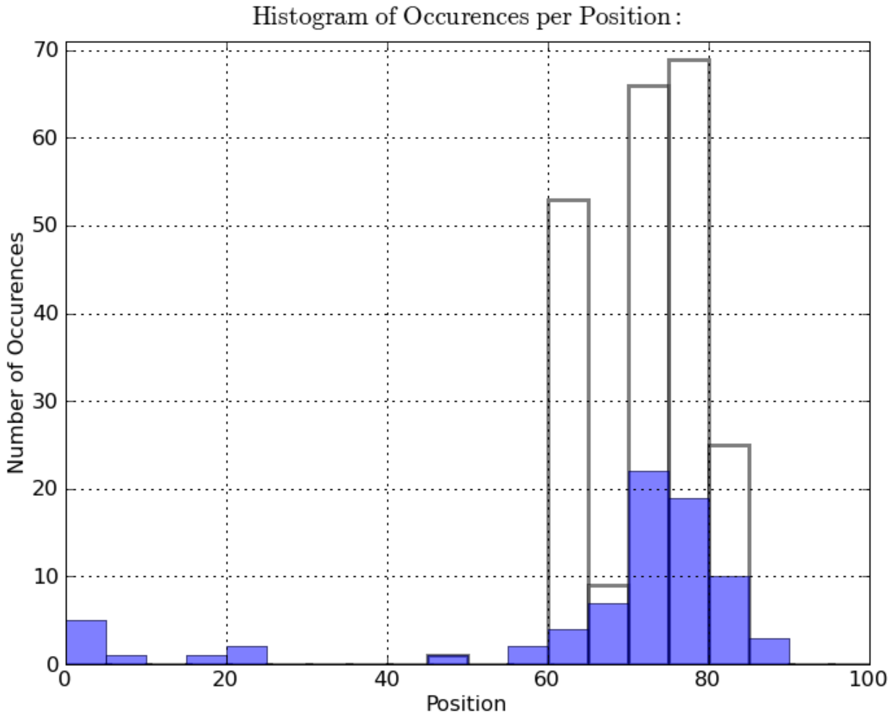


Figure 9: The occurrence frequency of triplet ICA at exo-loop 1 in amine subfamily (represented by white bars) and the other subfamilies (represented by blue).

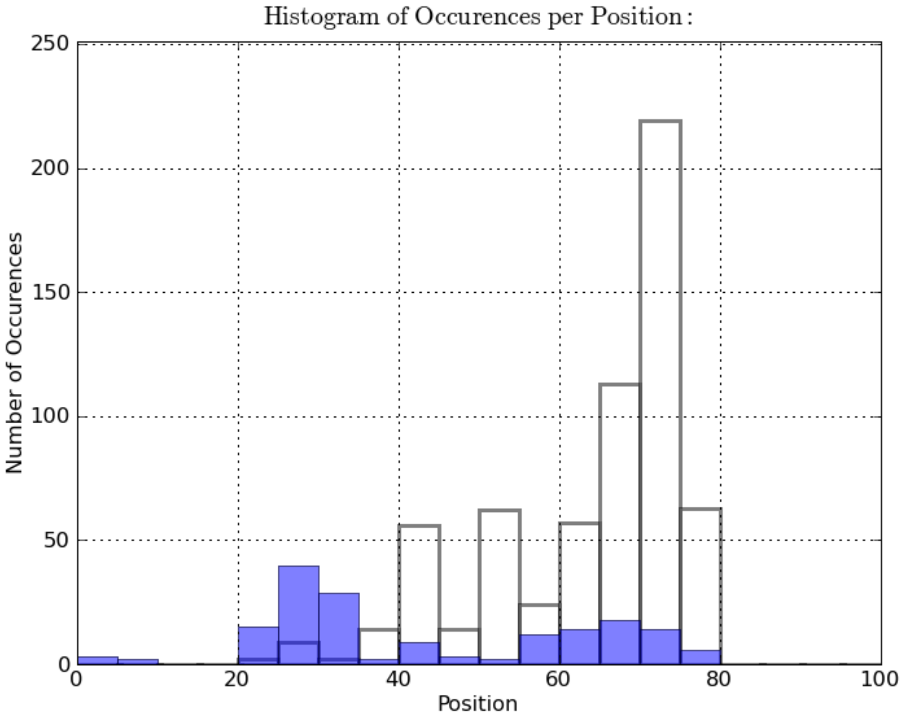


Figure 10: The occurrence frequency of triplet AIB at exo-loop 1 in peptide subfamily (represented by white bars) and the other subfamilies (represented by blue).

The performance analysis was performed on the subfamilies of Class A. Only the Class A family of GPCR sequences was used because of two reasons. First and foremost is that the GPCRpred dataset on which the classifier was tested contains the subfamily information for only the sequences in Class A. Secondly, more than 80% of the human GPCR sequences are grouped in this family; therefore, it is the most important target of pharmaceutical research.

It should be noted that the GPCRBind algorithm requires preprocessing of sequences by a trans-membrane prediction software (for which purpose TMHMM was used). For some sequences the TMHMM software did not predict a valid GPCR model. Therefore, those sequences for which TMHMM software can make an accurate prediction were used. This is a side-effect that has to be tolerated in order to discover the ligand interaction sites. As you can see from Table 3, no sequences are lost due to this reason for some of the subfamilies. For most of the remaining subfamilies, the amount of sequences that were eliminated in this manner are not significant. The only subfamily for which there was a significant drop in the number of sequences was the Prostanoid subfamily.

The GPCRBind method, proposed in this thesis, requires random partitioning. Due to this randomness the results of two successive runs are not identical. Therefore, the whole method is repeated 100 times and the average accuracy is reported.

The runtime of the algorithm versus the number of runs in the DPE step is shown in Figure 11. The runtime is linear with the number of runs at the DPE step as you can see in Figure 11. After the classification rules are generated, which is an offline step and performed only once. The classification

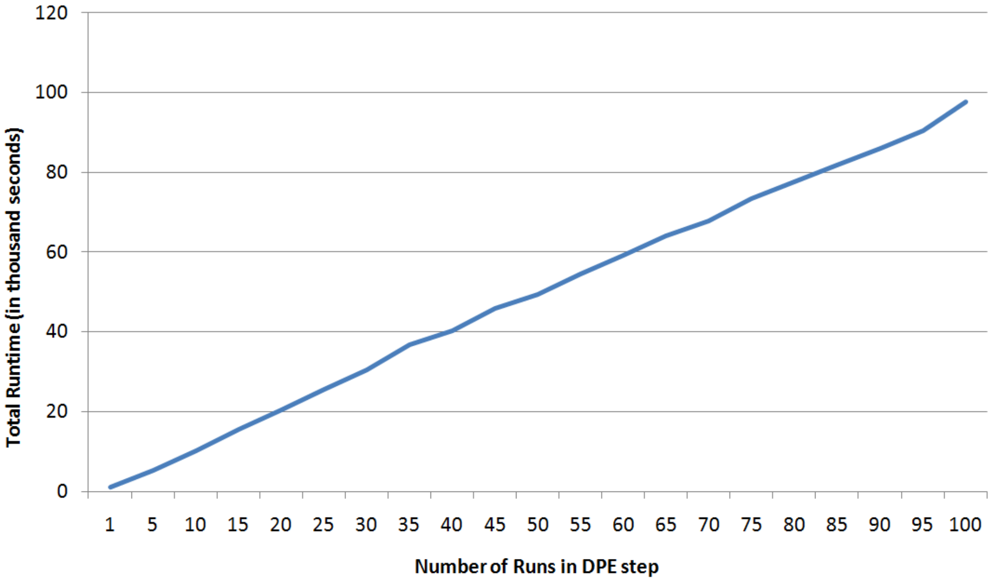


Figure 11: The runtime of the algorithm plotted against the number of runs in the DPE step with 70% DPE motif selection threshold.

takes less than a second to produce a classification for any given sequence.

5.2.1 Comparison with the GPCRpred Server

The performance of GPCRBind was compared against a recent GPCR classification server, GPCRpred, which predicts Class A subfamily membership information. In order to keep every factor constant during the testing of the two methods, the GPCRBind algorithm was trained with the GPCRpred dataset. The TMHMM-eliminated sequences were removed from the GPCRpred dataset and the remaining sequences were classified with both the GPCRpred server and the GPCRBind algorithm. Consequently the two techniques were trained and tested on exactly the same sequences.

For each sequence, GPCRpred first tries to predict if this sequence is a member of the GPCR superfamily or not whereas GPCRBind directly assumes that this sequence is a GPCR. The reason is that GPCRBind requires a priori determination of exo-cellular loops – which can only be achieved if the sequence already belongs to the GPCR set of sequences. This restriction of GPCRBind is due to its design as a discovery and exploration tool in addition to being a classification tool. However, this difference in the way the two classifiers work should create only a limited problem because the results reported in [6] claim that GPCRpred can distinguish a GPCR sequence from a non-GPCR with an accuracy of 99.5%. The detailed classification results corresponding to GPCRBind for individual subfamilies, taken from the best performing repetition out of 100 repetitions, is provided in Table 4. The averaged accuracy of 100 repetitions of GPCRBind is also shown at the bottom of this table. The number of runs used in the DPE step of GPCRBind is 20.

GPCRBind had a higher overall classification accuracy, but more importantly it had very high accuracy for all the subfamilies while GPCRpred performed poorly in some of the small-sized subfamilies. If the performance is evaluated solely based on overall accuracy, performance on large-sized subfamilies shadows classification quality on smaller-sized subfamilies: The confusion matrix of the best repetition of GPCRBind out of 100 repetitions is shown in Table 5. However, the fact that the number of sequences in a subfamily is small does not mean that it is insignificant. On the contrary, there is little correlation between the number of sequences in a subfamily and its significance to biotechnology research. Therefore, an ideal classification tool should perform equally well on both small-sized and large-sized subfamilies. GPCRBind performs extremely well on these small-sized subfamilies, achieving 100% classification performance for most of them whereas the SVM-based GPCRpred exhibits poor results such as 37.5% for prostanoid or 55.5% for gonadotrophin releasing hormone subfamilies.

It is evident that the DPE algorithm is very powerful in determining distinguishing motifs for every single subfamily. This also enhances the confidence in the ligand interaction sites discovered by this study. This knowledge is crucial for drug designers targeting GPCRs because it enables them to specifically target one subfamily but not the other.

5.2.2 Independent Dataset Comparison

To establish a new classification technique, an independent dataset testing is essential. Therefore, in the testing stage, GPCRBind was trained and tested on separate datasets. The training and testing datasets were chosen such that the results could be compared to state-of-the-art GPCR classification methods reported by Davies et al., in [1]. Davies et al. trained GPCRTree on the GDS dataset and then used GPCRTree to predict the subfamily of Class A sequences in the GPCRpred dataset. They compare their results to those given by PRED-GPCR on the same testing set. To be able to draw a direct comparison between GPCRTree's performance and that of the method, GPCRBind was also trained on GPCRTree's training set, namely the GDS dataset, and tested on the same GPCRpred dataset. As the DPE step involves randomness, the whole method has been repeated 100 times and the average accuracy over all the repetitions is presented. It can be seen from the results in Table 6 that GPCRBind performed superior to other classifiers when executed with 20 runs in the DPE step and a DPE motif

Subfamily	Number of sequences	Correctly processed by TMHMM
Amine(AMN)	221	208
Cannabinoid(CAN)	11	11
Gonadotrophin releasing hormone (GRH)	10	9
Hormone proteins(HMP)	25	24
Lysospingolipids(LYS)	9	8
Melatonin(MEL)	13	13
Nucleotide-like(NUC)	48	33
Olfactory(OLF)	87	69
Platelet activating factor (PAF)	4	4
Peptide(PEP)	381	304
Prostanoid(PRS)	38	8
Rhodopsin(RHD)	183	174
Thyrotropin releasing hormone (TRH)	7	7
Viral(VIR)	17	13
Total	1054	885 (84.0%)

Table 3: The number of sequences correctly processed by TMHMM in each subfamily.

Subfamily	Total	GPCRBind	GPCRpred
Peptide	304	302 (99.3%)	301 (99.0%)
Amine	208	203 (97.6%)	204 (98.1%)
Rhodopsin	174	169 (97.1%)	174 (100%)
Olfactory	69	68 (98.5%)	60 (86.9%)
Nucleotide-like	33	29 (87.8%)	24 (73.7%)
Hormone Protein	24	24 (100%)	21 (87.5%)
Viral	13	12 (92.3%)	0 (0%)
Melatonin	13	13 (100%)	10 (76.9%)
Cannabinoid	11	9 (81.8%)	9 (81.8%)
GRH	9	9 (100%)	5 (55.5%)
Prostanoid	8	8 (100%)	3 (37.5%)
Lysospingolipids	8	8 (100%)	6 (75%)
TRH	7	6 (85.7%)	4 (57.1%)
PAF	4	1 (25.0%)	0 (0%)
Overall	885	861 (97.3%)	821 (92.8%)
100 Repetitions	885	851.3 (96.2%)	821 (92.8%)

Table 4: Classification performance of GPCRBind and GPCRpred.

		Actual													
		PEP	AMN	RHD	OLF	NUC	HMP	VIR	MEL	CAN	GRH	PRS	LYS	TRH	PAF
Predicted	PEP	302	5	4	1	3	0	0	0	2	0	0	0	1	3
	AMN	1	203	1	0	1	0	1	0	0	0	0	0	0	0
	RHD	1	0	169	0	0	0	0	0	0	0	0	0	0	0
	OLF	0	0	0	68	0	0	0	0	0	0	0	0	0	0
	NUC	0	0	0	0	29	0	0	0	0	0	0	0	0	0
	HMP	0	0	0	0	0	24	0	0	0	0	0	0	0	0
	VIR	0	0	0	0	0	0	12	0	0	0	0	0	0	0
	MEL	0	0	0	0	0	0	0	13	0	0	0	0	0	0
	CAN	0	0	0	0	0	0	0	0	9	0	0	0	0	0
	GRH	0	0	0	0	0	0	0	0	0	9	0	0	0	0
	PRS	0	0	0	0	0	0	0	0	0	0	8	0	0	0
	LYS	1	0	0	0	0	0	0	0	0	0	0	8	0	0
	TRH	0	0	0	0	0	0	0	0	0	0	0	0	6	0
	PAF	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 5: Confusion matrix of GPCRBind on GPCRpred dataset.

Classifier	Accuracy
GPCRBind	90.7%
GPCRTree	76.2%
PRED-GPCR	73.8%

Table 6: Classification accuracy of GPCRBind compared to the results reported by Davies et al. [1].

selection threshold of 70%. In Table 6, the classification accuracy reported for GPCRBind is the averaged result of 100 repetitions to smooth out the effect of the randomness in the DPE step. It should be noted that all of the 100 repetitions of GPCRBind yielded results that are superior to the previous classifiers.

5.3 Classification Results of Sub-subfamilies of Amine Subfamily

The GPCRBind has been used to classify the sub-subfamilies of the Amine subfamily to have a better picture about its ability to be a general GPCR classifier. The GPCR sequences of the sub-subfamilies of the Amine subfamily have been retrieved from GPCRDB [29]in July 2010. In order to effectively mine rules about the potential ligand-receptor interaction sites, the algorithm has been trained on the whole data and consequently trained on the entire dataset. Table 7 shows the number of sequences in each sub-subfamily in the original dataset retrieved from GPCRDB compared to the number of sequences left after the TMHMM processing.

In an effort to improve the classification performance on the sub-subfamilies a number of changes were implemented. The first of these is to reduce the

Subfamily	Number of sequences	Correctly processed by TMHMM
Adrenoreceptors (ADR)	483	304
Dopamine (DOP)	317	240
Histamine (HIS)	183	136
Muscarinic Acetylcholine (MUS)	198	169
Octopamine (OCT)	91	66
Serotonin (SER)	575	477
Trace Amine (TRA)	257	216
Total	2104	1608 (76.4%)

Table 7: The number of sequences correctly processed by TMHMM in each Amine sub-subfamily.

number of positions from 10 to 3. This was done with the intuition that 10 positions provided unnecessarily detailed positional information. Secondly, in an effort to increase the runtime of the algorithm and improve the quality of the findings, the absence conditions in each rule retrieved from the decision tree were removed automatically. After the rule extraction algorithm extracts rules from the decision tree, there are a huge number of absence conditions in each rule. This is because a decision tree can only classify one class at each node as it can only test for a single motif at any node. Therefore classifying the rules extracted from the lower levels include a high number of absence rules that do not necessarily contribute to the classification but instead that are simply relics of the classification decisions made in the higher levels of the decision tree. Therefore most of these absence rules get eliminated by the condition filtering step of the rule extraction method that uses Fisher's exact test to measure the contribution of each condition to the classification effort. The few absence conditions that are left in the rule body actually reduce the potential contribution of the findings to drug design efforts - drug designers are more interested in the motifs that appear in the protein than those that are absent. Therefore elimination of the absence conditions in a preprocessing step of each rule, both saves computational time and improves the contribution of the results to drug designers. The last change made was to convert the DPE motif selection threshold to a fixed integer value, denoting the number of motifs with the highest DPE score to pick for use in the rule extraction step.

The performance of GPCRBind with the above-mentioned improvements has been tested on the sub-subfamilies of the Amine subfamily with a DPE motif selection threshold of 500 - i.e. top 500 motifs with respect to the DPE scores have been selected for rule extraction. The accuracy of the classification is 81.3% and the corresponding confusion matrix is given in Table 8.

5.4 Accuracy-Runtime Trade-off

Two factors determine the runtime of GPCRBind: the number of runs in the DPE step, and the DPE motif selection threshold. The number of runs in the DPE step has a linear impact on the runtime. It was observed that after 20 runs, there is marginal contribution to the classification performance. Therefore, 20 runs were used for all the presented results. On the contrary, evaluating the effect of the DPE motif selection threshold on the runtime is non-trivial. DPE motif selection threshold aims to eliminate motifs that have

		Actual						
		ADR	DOP	HIS	MUS	OCT	SER	TRA
Predicted	ADR	250	17	4	5	5	9	2
	DOP	4	165	4	0	0	14	1
	HIS	2	2	79	3	3	4	1
	MUS	4	1	10	141	0	23	0
	OCT	0	7	0	0	57	0	1
	SER	44	48	35	16	4	424	20
	TRA	0	0	4	4	4	3	191

Table 8: Confusion matrix of GPCRBind on the Amine sub-subfamilies.

low distinguishing power. The higher this threshold, the higher the number of motifs selected for use in rule extraction. A larger number of motifs means a higher number of attributes for each sequence, thus contributing more information. However, an increase in the number of attributes exponentially increases the complexity of the rule extraction process. This non-linear and complex relationship has been investigated by alternating the DPE motif selection threshold while performing independent dataset classification. Figure 12 shows the accuracy-runtime trade-off for different threshold values. In the figure, 100 repetitions of the whole method have been performed and the average accuracy and runtime is reported for every threshold value. As can be seen in the figure, accuracy rises sharply as the threshold goes from 80% to 75% and from 75% to 70%. However, as the threshold goes from 70% to 60%, there is only a slight increase in accuracy at the cost of a significant increase in runtime. Therefore, it was concluded that 70% is the optimum threshold value in terms of runtime-accuracy trade-off for training on the GDS dataset.

It should be noted that GPCRBind is a rule extraction method, and while training takes time on the order of hours, classification of a sequence takes milliseconds. This property of GPCRBind makes it suitable for being used as a classification server.

5.5 Interaction Site Discovery Results

Each classification rule used by GPCRBind is a multitude of conditions regarding the presence/absence of motifs which, if satisfied, claims that the sequence belongs to a particular subfamily with a specific certainty factor.

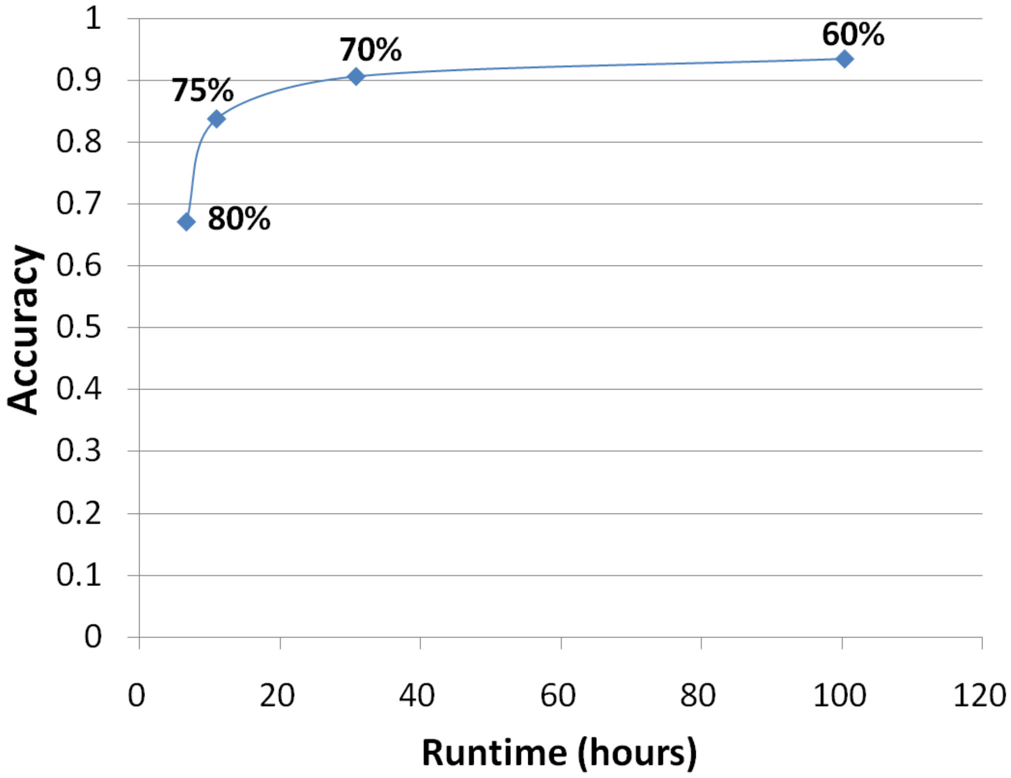


Figure 12: The accuracy versus runtime for the following DPE motif selection threshold values: 60%, 70%, 75% and 80%.

Each motif, is a triplet occurrence at a specific site. Therefore, in essence, GPCRBind rules predict subfamily membership based on the presence/absence requirement of certain amino acids, each at a particular position. Rules that characterize each subfamily by amino acid presence/absence rules at their exo-cellular loops have been discovered through this research. The entire set of rules is given in Table 9, which shows that GPCRBind can successfully distinguish GPCR subfamilies with only a few motifs for each subfamily. In Table 9 for every rule, the following information is given: the triplet (in Sezerman encoding), the region of occurrence and the position of occurrence within that region. Whether if the condition is for presence or absence of the motif is also indicated. For selected subfamilies, the rule with the highest certainty factor on Table 9 is represented on a GPCR snake-diagram in Figure 13 which visualizes the findings of the method. In this figure, the rule with the highest CF score on Table 9 has been represented for 5 subfamilies: AMN, HMP, PRS, RHD and TRH. Boxes ($\diamond$) represent the location of the motifs as shown on Table 9, and the size of the box is proportional to the positions that the motif spans within that region. The initial letter of the subfamily name is placed in the box corresponding to the motifs of that subfamily. In cases where a rule is composed of two motif presence conditions, one box is shown for each presence condition.

In summary this thesis has two novel contributions; a powerful classification technique and a way to predict interaction sites of GPCRs from sequence information alone.

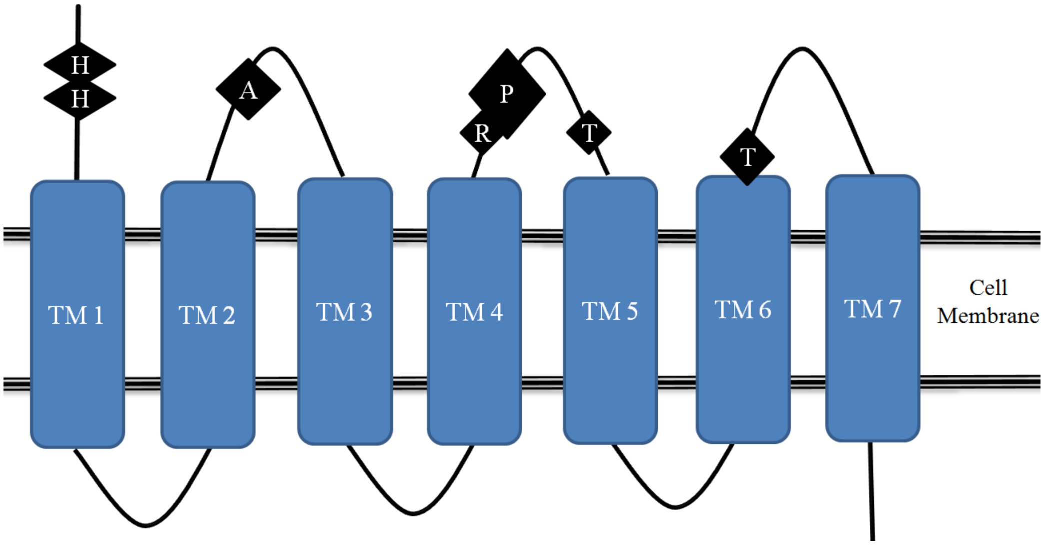


Figure 13: Representation of rule conditions on a GPCR snake-diagram.

6 Conclusions and Future Work

In this thesis a technique for GPCR classification through the discovery of key ligand interaction sites was proposed. The proposed methods were implemented and tested on real datasets. The results are compared with the state-of-the-art GPCR classification techniques. Experiments show that GPCRBind outperforms the state-of-the-art classification techniques.

GPCRBind is planned to be developed by generalizing and applying it to broader subfamilies of receptors. The only limitation for applying GPCRBind to other receptors is the necessity of having well-defined region information in the sequence. When a new and more general motif definition that can work with or in the absence of well-defined regions is developed, the DPE step and the remaining part of the algorithm can be directly applied. The next step is to develop new motif definitions to accomplish this.

Subfamily	CF	Triplet	Presence/ Absence	Region	Position
TRH	0.79	JAA	1	ECL2	8,9
		AAA	1	ECL3	0,1
PAF	0.88	EED	1	ECL2	0,1,2,4
		JEA	1	NTERM	7
LYS	0.83	EIE	1	ECL2	3,4
		AKA	1	ECL2	6,8
MEL	0.96	JCK	1	ECL2	1
		CEA	0	ECL3	0,1,6
AMN	0.98	HKA	1	ECL1	3,4
	0.97	EBA	1	ECL1	1,5,6
HMP	0.98	JAA	1	NTERM	3,4,5
		EGA	1	NTERM	3,6,7
NUC	0.94	GJI	1	ECL1	1,2
	0.94	EAG	1	ECL2	3
VIR	0.75	JCK	1	ECL2	1
		CEA	1	ECL3	0,1,6
	0.85	ABC	1	ECL3	3,7,8
PRS	0.94	EHI	1	ECL2	2,3,4
GRH	0.61	DAE	1	ECL1	0
		AJI	0	ECL1	7

Table 9: Selected rules for each subfamily.

DPE algorithm might possibly be improved by substituting alternatives to the ID3 algorithm. One of the first alternatives to ID3 is the J48 algorithm which is a derivative of C4.5 and generally considered to be an advanced decision tree induction algorithm. However, within the DPE algorithm, the main object of using a classifier is to detect the most distinguishing motif. Instead of using a decision tree algorithm, another motif evaluation scheme might be introduced as well. One way to accomplish this is by formulating a "fitness score" and running an optimization algorithm to increase the fitness as much as possible. Another alternative is to interpret the motif presence as a coverage function and find the optimal coverage for the training set. These changes to the DPE algorithm might potentially result in an improved performance.

To improve the GPCRBind technique, alternative rule extraction methods can be tested in the future. The Particle Swarm Optimization / Ant Colony Optimization (PSO/ACO) rule extraction algorithm described in [19] is an alternative that deserves future testing. The motifs used in the referred work were picked from other pre-existing resources. It is possible that the motifs discovered by the DPE algorithm might serve the PSO/ACO algorithm better and therefore result in a higher accuracy.

This thesis contributes a novel motif evaluation metric, MSM, to the GPCR classification effort. MSM can be generalized for use with motifs from a whole range of domains to evaluate their performance with respect to a given classification problem. As long as the central dogma of modern biology (which states that sequence plays a significant role in the function of a protein) stands, motif extraction/evaluation methods are required. Similarly, the DPE algorithm does not simply provide a solution to this particular problem, but instead it can be applied in many other situations where motif extraction is required. Therefore this thesis prepares the framework for these motif extraction/evaluation algorithms to be used in a multitude of other research projects.

GPCRBind operates on the basis that the extra-cellular regions of the protein is known. This requires *a priori* parsing of the sequence to identify the extra-cellular regions. In the future, it needs to be seen whether if extracting motifs from the other regions of the protein contribute to classification accuracy. A number of small preliminary experiments done in this direction have shown decreased classification accuracy, which has led to abandoning the idea at a rather early stage. However, testing with different combinations of motifs formed from a wider range of regions and differentiated classifiers might potentially yield better classification performance.

GPCRBind can easily be adapted as a web-based classification server. Given a query sequence, TMHMM (or another trans-membrane region prediction tool) is used to identify the trans-membrane helices. If the tool can correctly identify a 7TM structure, GPCRBind is invoked and classification is performed. Otherwise, the query sequence is classified as a non-GPCR. Therefore the GPCRBind technique proposed here holds great potential to satisfy the need for well-performing GPCR classifiers.

GPCRBind is a GPCR classification specific technique taking into account domain knowledge while existing classification servers employ very general classification tools such as SVM or HMM which are designed to classify any type of data. The existing methods require large training sets to successfully learn small-sized subfamilies whereas GPCRBind can effectively learn from a few sequences. Consequently the performance of the other classifiers approach the performance of the proposed technique only for subfamilies with many sequences. However, from a drug design perspective, the importance of a subfamily is not always correlated with the number of sequences within that subfamily. As it takes advantage of problem specific information, GPCRBind is more successful for this classification problem and also more helpful to biomedical researchers.

The results of the classification of the Amine sub-subfamilies indicates that the technique is more successful at distinguishing subfamilies. One reason behind this fact is that the ligands of the receptors within the same subfamily are significantly more similar compared to the ligands of the receptors in the same family. As the ligands are more similar, it is only natural to expect that the sites that recruit or bind those ligands are more similar as well. The sub-subfamily classification problem needs to be studied in greater detail in order to create a top-down, hierarchical classifier.

The subfamily characterization produced by this study is very successful in distinguishing members of one subfamily from another – as shown in Sections 5.2.1 and 5.2.2. The most plausible explanation for variation in the exo-cellular regions of the sequences in two different subfamilies is the difference dictated by the physico-chemical requirements of binding to their respective ligands. To illustrate, the exo-cellular variation between peptide and amine binding GPCR sequences can only be attributed to the different physico-chemical properties required for binding to peptides or amines. Therefore, the rules that GPCRBind discovers are essentially interaction sites between ligands and receptors in a subfamily and the ligand set of that subfamily. If these sites were not ligand-specific then GPCRBind would not have been able to distinguish members of each subfamily with high accuracy using these rules. If these sites are indeed ligand-specific, given that they are exo-cellular, there is a very strong chance that they play a major role in receptor agonism. Since very few GPCR structures are known and very few ligand receptor binding studies are carried out experimentally, these sites might help to cope with the difficulty in discovering subfamily-selective drug candidates to pharmaceutical researchers.

References

[1] M. Davies, A. Secker, A. Freitas, M. Mendao, J. Timmis, and D. Flower, “On the hierarchical classification of G protein-coupled receptors,” *Bioinformatics*, vol. 23, no. 23, p. 3113, 2007.

[2] D. Cojocari. (2010, Jul.) Amino acids. [Online]. Available: http://upload.wikimedia.org/wikipedia/commons/0/0f/Amino_acids.png

[3] D. Filmore, “It is a GPCR world,” *Modern drug discovery*, vol. 7, no. 11, pp. 24–28, 2004.

[4] T. Hebert and M. Bouvier, “Structural and functional aspects of G protein-coupled receptor oligomerization,” *Biochemistry and Cell Biology*, vol. 76, no. 1, pp. 1–11, 1998.

[5] T. Kenakin, “Allosteric modulators: the new generation of receptor antagonist,” *Molecular interventions*, vol. 4, no. 4, p. 222, 2004.

[6] M. Bhasin and G. Raghava, “GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors,” *Nucleic Acids Research*, vol. 32, no. Web Server Issue, p. W383, 2004.

[7] R. Karchin, K. Karplus, and D. Haussler, “Classifying G-protein coupled receptors with support vector machines,” *Bioinformatics*, vol. 18, no. 1, p. 147, 2002.

[8] B. Bakir and O. Sezerman, “Functional Classification of G-Protein Coupled Receptors, Based on Their Specific Ligand Coupling Patterns,” *Applications of Evolutionary Computing*, pp. 1–12, 2006.

[9] P. Papasaikas, P. Bagos, Z. Litou, V. Promponas, and S. Hamodrakas, “PRED-GPCR: GPCR recognition and family classification server,” *Nucleic acids research*, vol. 32, no. Web Server Issue, p. W380, 2004.

[10] J. Cui, L. Han, H. Li, C. Ung, Z. Tang, C. Zheng, Z. Cao, and Y. Chen, “Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties,” *Molecular immunology*, vol. 44, no. 4, pp. 514–520, 2007.

[11] A. Secker, M. Davies, A. Freitas, E. Clark, J. Timmis, and D. Flower, “Hierarchical classification of G-Protein-Coupled Receptors with data-driven selection of attributes and classifiers,” *International journal of data mining and bioinformatics*, vol. 4, no. 2, pp. 191–210, 2010.

[12] W. Atchley, J. Zhao, A. Fernandes, and T. Dr
"uke, “Solving the protein sequence metric problem,” *Proceedings of the National Academy of Sciences*, vol. 102, no. 18, p. 6395, 2005.

[13] P. Joost and A. Methner, “Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands,” *Genome Biol*, vol. 3, no. 11, pp. 1–16, 2002.

[14] K. Chou and D. Elrod, “Bioinformatical analysis of G-protein-coupled receptors,” *Journal of Proteome Research*, vol. 1, no. 5, pp. 429–433, 2002.

[15] K. Chou, “Prediction of G-protein-coupled receptor classes,” *J. Proteome Res*, vol. 4, no. 4, pp. 1413–1418, 2005.

[16] T. Attwood, “A compendium of specific motifs for diagnosing GPCR subtypes,” *Trends in pharmacological sciences*, vol. 22, no. 4, pp. 162–165, 2001.

[17] D. Flower and T. Attwood, “Integrative bioinformatics for functional genome annotation: trawling for G protein-coupled receptors,” in *Seminars in cell & developmental biology*, vol. 15, no. 6. Elsevier, 2004, pp. 693–701.

[18] N. Mulder, R. Apweiler, T. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, V. Buillard, L. Cerutti, R. Copley *et al.*, “New developments in the InterPro database,” *Nucleic Acids Research*, vol. 35, no. Database issue, pp. D224–D228, 2007.

[19] N. Holden and A. Freitas, “Hierarchical classification of G-protein-coupled receptors with a PSO/ACO algorithm,” in *Proceedings of the IEEE Swarm Intelligence Symposium (SIS’06).* IEEE Press, 2006, pp. 77–84.

[20] M. Davies, D. Gloriam, A. Secker, A. Freitas, M. Mendao, J. Timmis, and D. Flower, “Proteomic applications of automated GPCR classification,” *Proteomics*, vol. 7, no. 16, pp. 2800–2814, 2007.

[21] Y. Yabuki, T. Muramatsu, T. Hirokawa, H. Mukai, and M. Suwa, “GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model,” *Nucleic acids research*, vol. 33, no. Web Server Issue, p. W148, 2005.

[22] U. Gether, “Uncovering molecular mechanisms involved in activation of G protein-coupled receptors,” *Endocrine reviews*, vol. 21, no. 1, p. 90, 2000.

[23] D. Rosenbaum, S. Rasmussen, and B. Kobilka, “The structure and function of G-protein-coupled receptors,” *Nature*, vol. 459, no. 7245, pp. 356–363, 2009.

[24] S. Foord, T. Bonner, R. Neubig, E. Rosser, J. Pin, A. Davenport, M. Spedding, and A. Harmar, “International Union of Pharmacology. XLVI. G protein-coupled receptor list,” *Pharmacological reviews*, vol. 57, no. 2, p. 279, 2005.

[25] M. Davies, A. Secker, M. Halling-Brown, D. Moss, A. Freitas, J. Timmis, E. Clark, and D. Flower, “GPCRTree: online hierarchical classification of GPCR function,” *BMC Research Notes*, vol. 1, no. 1, p. 67, 2008.

[26] P. Conn, A. Christopoulos, and C. Lindsley, “Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders,” *Nature Reviews Drug Discovery*, vol. 8, no. 1, pp. 41–54, 2009.

[27] F. Libert, M. Parmentier, A. Lefort, C. Dinsart, J. Van Sande, C. Maenhaut, M. Simons, J. Dumont, and G. Vassart, “Selective amplification and cloning of four new members of the G protein-coupled receptor family,” *Science*, vol. 244, no. 4904, p. 569, 1989.

[28] A. Methner, G. Hermey, B. Schinke, and I. Hermans-Borgmeyer, “A novel G protein-coupled receptor with homology to neuropeptide and

chemoattractant receptors expressed during bone development," *Biochemical and biophysical research communications*, vol. 233, no. 2, pp. 336–342, 1997.

[29] F. Horn, J. Weare, M. Beukers, S. Horsch, A. Bairoch, W. Chen, O. Edvardsen, F. Campagne, and G. Vriend, "GPCRDB: an information system for G protein-coupled receptors," *Nucleic acids research*, vol. 26, no. 1, p. 275, 1998.

[30] A. Krogh, B. Larsson, G. Von Heijne, and E. Sonnhammer, "Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1," *Journal of molecular biology*, vol. 305, no. 3, pp. 567–580, 2001.

[31] M. Davies, A. Secker, A. Freitas, E. Clark, J. Timmis, and D. Flower, "Optimizing amino acid groupings for GPCR classification," *Bioinformatics*, vol. 24, no. 18, p. 1980, 2008.

[32] B. Erguner, O. Erdogan, and U. Sezerman, "Prediction and classification for GPCR sequences based on ligand specific features," *Computer and Information Sciences–ISCIS 2006*, pp. 174–181, 2006.

[33] S. Moller, M. Croning, and R. Apweiler, "Evaluation of methods for the prediction of membrane spanning regions," *Bioinformatics*, vol. 17, no. 7, p. 646, 2001.

[34] G. Salton, "Developments in automatic text retrieval," *Science*, vol. 253, no. 5023, pp. 974–980, 1991.

[35] J. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[36] ——, “Simplifying decision trees,” *International journal of man-machine studies*, vol. 27, no. 3, pp. 221–234, 1987.