Analysis of the Impact of Transcript Diversity on Protein Domains of G-Protein-Coupled Receptors (GPCRs) in Human, Mouse and Rat Proteomes: A Data Mining Approach

Felix Olanrewaju Babalola

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Engineering

Eastern Mediterranean University
June 2017
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Mustafa Tümer
Director

I certify that this thesis satisfies the requirements as a thesis for the degree of Master of Science in Computer Engineering.

Prof. Dr. H. Işık Aybay
Chair, Department of Computer Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Computer Engineering.

Assoc. Prof. Dr. Bahar Taneri
Co-Supervisor

Assoc. Prof. Dr. Ekrem Varoğlu
Supervisor

Examining Committee

1. Prof. Dr. Hakan Altınçay

2. Prof. Dr. Doğu Arifler

3. Assoc. Prof. Dr. Bahar Taneri

4. Assoc. Prof. Dr. Ekrem Varoğlu

5. Asst. Prof. Dr. Adil Şeytanoğlu

ABSTRACT

The modern medicine industry and other related industries have become more interested in the structures and functionalities of G-protein coupled receptors (GPCRs) because researches have shown that a good number of drugs act by binding to GPCRs in human and other close mammalian organisms. Motivated by this, this study analyzed three genomes; human, mouse and rat, for the presence and the extent of protein domain diversity. The aim is to provide a direction to other researches, on how and how much drugs bind to GPCRs by confirming the presence of differences in protein domains coded by transcripts of genes.

Public biological databases with comprehensive datasets about various genomes and proteomes were used in this study. Data relevant to this study were retrieved from preferred biological databases. These are then stored in a separate database created for this study and analyzed based on this study.

Results of our analysis showed that differences exist in GPCR protein domain in all three genomes, and that this is influenced by transcript diversity. It was found that for human, 83 percent of GPCR genes with multiple transcripts exhibits diversity in the domains they code for. This was found to be 81 and 65 percent for mouse and rat respectively. This implies that further study on factors leading to these diversities could go a long way in helping to identify structures, mutations and functions of GPCRs and consequently would be of benefit to drug development and related studies.

Keywords: G-protein coupled receptors, genomes, proteomes, transcript, protein domain, biological databases, data retrieval, data storage and analysis.

ÖZ

G-proteine bağlı reseptörlerin (GPCR) yapı ve işlevsellikleri modern tıp ve alakalı endüstrileride büyük ilgi görmektedir; çünkü araştırmalar göstermiştir ki insanlarda ve birçok memelide çeşitli ilaçlar GPCRlere bağlanarak çalışmaktadır. Bu bilgiden hareketle, bu çalışmada insan, fare ve sıçan genomları analiz edilmiş ve bu organizmaların protein çeşitliliği araştırılmıştır. Buradaki amaç, GPCR genlerinin transkriptlerinin kodladığı proteinlerdeki farklılıkları tanımlayarak, GPCR ilaç etkileşimi alanındaki araştırmalara katkı sağlamaktır.

Bu çalışmada, kamuya açık biyolojik veritabanlarında yer alan genom ve proteomlarla ilgili detaylı veri setleri kullanılmıştır. İlgili veriler gerekli veritabanlarından alınmıştır. Daha sonra bu veriler bu çalışma için ayrıca yaratılan bir veritabanına aktarılmış ve analiz edilmiştir.

Çalışma sonucunda elden edilen analiz sonuçları, her üç genomda da GPCR protein çeşitiliği olduğunu göstermekte ve bu çeşitliliğin transkript çeşitliliğinden kaynaklandığını göstermektedir. İnsan genomundaki GPCR genlerinden birden çok transkripti olanların, yüzde 83ünde protein çeşitliliği saptanmıştır. Bu oran farede yüzde 81 olup, sıçanda ise yüzde 65tir. Bu sonuçların sebepleri ileri çalışmalar ile aydınlatılabilir ve böylece GPCRlerin yapı, fonksiyon ve mutasyonları daha iyi anlaşılıp, ilaç geliştirme alanında fayda sağlanabilir.

Anahtar Kelimeler: G-proteine bağlı reseptörler, genomlar, proteomlar, transkript, biyolojik veritabanları, veri çıkarımı, veri saklama ve analiz.

To my family and friends

ACKNOWLEDGMENT

My foremost gratitude is to Assoc. Prof. Dr. Ekrem Varoğlu, my supervisor, and Prof. Dr. Bahar Taneri, my co-supervisor; I could not have completed this work without their guidance, motivation and ideas. They treated me as part of a team, as a friend, creating a working environment better than any student could ever ask for. Their encouragement triggered more interest in bioinformatics and related research in me.

I will also like to express my sincere gratitude to Prof. Dr.H. Işık Aybay, the Chair of Department of Computer Engineering and all the faulty members and staffs of the Department for their support. Working as a Research Assistant in this Department was of great help for the period that I worked on this thesis.

Special gratitude to my brother, Stephen Babalola for his support, he is the reason I made it thus far. And to my parents and my other siblings, they never stopped believing in me, I'm very grateful to them all.

Many thanks also to friends who encouraged and supported me during this study especially members of Advisory Board of St Cyril's Catholic Community, Damilola Aransiola and Abolade Osanyintoba.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Background

G-protein-coupled receptors (GPCRs) are special receptors which received the attention of researchers over the years, with a lot of work being done to understand their structures and functions. The modern pharmaceutical industry is one particular industry that has been heavily interested in the interaction of these receptors with certain enzymes and drugs because research has shown that about 33% to 50% of drugs act by interacting with GPCRs present in human and other organisms. Understanding the mechanism of action of GPCRs as they make contact with other components of the body is therefore of great importance in the production and/or enhancement of drugs for better efficiency.

There are numerous GPCRs genes (about a thousand of them) and G proteins that they bind to. Each of these genes has different transcript(s), mostly more than one; this translates into differences in protein structures. In specific, domain differences arise, which also translates into functional differences. It is therefore important to study the transcript diversity of these genes and document the differences that exist between domains of proteins produced by each gene. This is the main research topic presented in this thesis.

Large data about GPCRs exist in various databases which will help us to analyze this topic. Analyzing and interpreting these data from various databases including protein domains, protein structures, nucleotide and amino acid sequences, require the development and implementation of tools that enhances efficient access to and management of different types of data. The goal of this work is to find computational and analytical solutions to these problems using appropriate tools and programming languages.

1.2 Thesis Contribution

The work described in this thesis analyzed the relationship between protein domain diversity in transcripts, their complexity and functionality in GPCRs. Three different genome namely human, mouse and rat, are analyzed individually for their transcript and protein domain diversity. The percentage of this diversity across all GPCRs as well as how this affects changes and complexity in the domains is analyzed and then the results from the three species are compared.

1.3 Thesis Outline

This chapter introduces the concept of this thesis; it contains the motivation behind this thesis. Chapter 2 gives an overview of molecular biology, concentrating on gene expression, while also giving information on the evolution of bioinformatics field and its usefulness in biological research. Chapter 3 introduces and gives details about GPCRs; discussing their structure and function. In Chapter 4, the methodology used to search, retrieve, store and analyze data used is highlighted, while Chapter 5 provides detailed explanation and illustration of results, as well as deductions from those results. Chapter 6 contains conclusion and proposed future work based on this thesis.

Chapter 2

OVERVIEW OF MOLECULAR BIOLOGY AND BIOINFORMATICS

2.1 Gene Expression

Gene expression is the process by which genetic information encoded in a gene is used to synthesize functional products. These products are usually proteins that are involved in essential activities in organisms as enzymes, hormones or as receptors [1]. Mainly, a protein product is produced through an initial step of RNA synthesis, referred to as transcription. This is then followed by protein synthesis, referred to as translation [2].

Genes are subunits of DNA, which is where information of a cell is stored. There are $3x10^9$ base pairs of DNA in every cell in the nucleus in humans which are distributed over 23 pairs of chromosomes and each cell has two copies of genetic materials which form the human genome. The human genome has about 20,310 genes, each coding particular protein(s) although about 95% of the genome is non-coding [3].

Figure 2.1 shows the central dogma of molecular biology, where DNA in an organism is the basic source of information. DNA is transcribed into RNA, and then RNA is translated into proteins, and DNA is continuously replicated to preserve itself and thereby able to pass on genetic information onto newly formed cells. Gene

expression is the combination of the processes of transcription and translation; they are further discussed in detail in sub-sections 2.1.1 and 2.1.2 respectively.

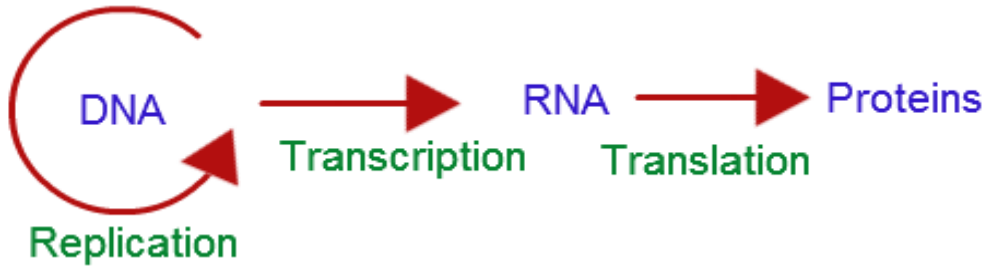


Figure 2.1: Central Dogma of Molecular Biology showing DNA as the basic origin for information in organisms (Figure taken from [4])

Gene expression is different across cells because some genes get transcribed while others do not. Every single cell has exactly the same DNA in them, but the cells have different functions which come about because of differential gene expression and consequently lead to cell specialization. These differences occur at development stage of cells just as regulatory mechanisms switch on and off.

DNA is usually wrapped around histone proteins in a structure called nucleosome. Loosely packed nucleosome is called **euchromatin** while tightly packed nucleosome is **heterochromatin.** Figure 2.2 shows the tightly packed appearance of heterochromatin and repetitive DNA sequences which all make it less transcriptionally active, against euchromatin, where genes are loosely packed with unique DNA sequences and transcriptionally active as a result of this structure.

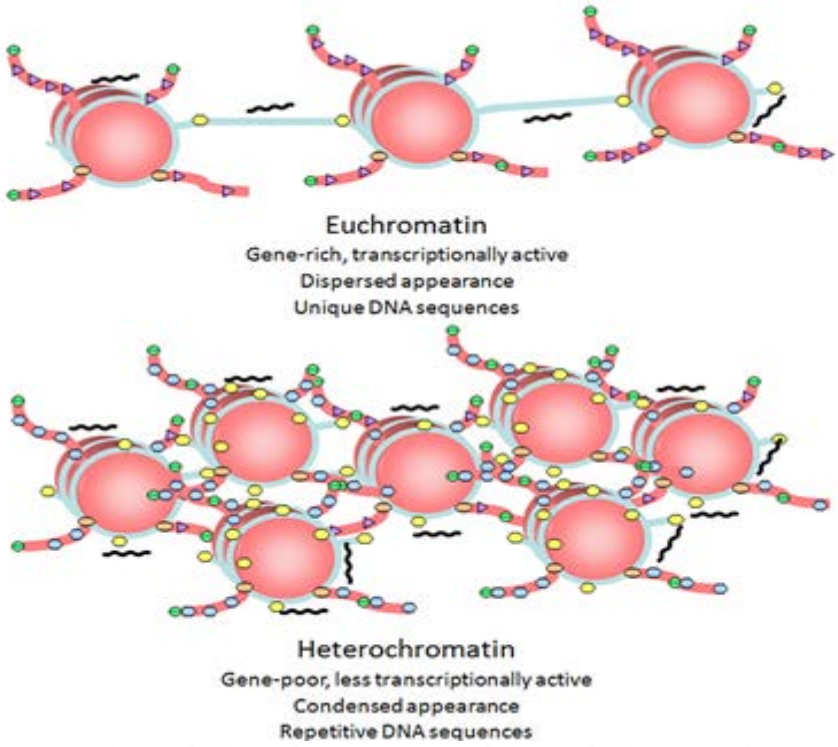


Figure 2.2: Loosely packed Euchromatin vs tightly packed Heterochromatin (Figure taken from [5])

2.1.1 Transcription

Transcription process occur when a strand of DNA, which stores genetic materials in the nuclei of cells, are copied into messenger RNA (simply called mRNA). mRNA is a molecule that is comparable to a copy of DNA, containing the same information. However, although they contain the same details, DNA and mRNA are not identical, as further discussed in Section 2.3

mRNA moves details of genetic materials contained in DNA from the nucleus to the ribosome; this forms the beginning of protein synthesis. Transcription is important because it produces mRNA strand necessary for translation.

RNA Polymerase is the enzyme needed for transcription as well as some accessory proteins known as **transcription factors**, together form transcription initiation

complex. The transcription factors attach to enhancer and promoter sequences in the DNA to trigger RNA polymerase to a transcription site. RNA polymerase matches complementary bases to the initial DNA strand to start the process of mRNA synthesis.

Transcription factors (TF) facilitate binding of RNA polymerase to DNA regions called **promoters,** which are, regulatory sequences that control transcription**.** Transcription starts at the promoters. However, some TFs are activators while some are repressors, and how much gene product will be made depend on specific combination of TFs. Signal transmission within and between cells mediates to activate TFs and therefore mediates gene expression. For example, cytokines and other growth factors regulate gene expression, aiding cell replication and division [5].

There are three types of RNA Polymerases (RNA Pol) in eukaryotic cells. RNA Pol I encodes a copy of the genes that encrypt most of the ribosomal RNAs, RNA Pol II encodes the messenger RNAs which is the most important component for protein molecules production, while RNA Pol III rewrites transfer RNAs (tRNAs) which are needed in the translation process, as well as other small regulatory RNA molecules [5].

Transcription process involves 2 steps, pre-messenger RNA (pre-mRNA) is first formed with the aid of RNA Pol enzymes, relying on Watson-Crick base pairing. The second step is RNA splicing involving reshaping the pre-mRNA to form the mature mRNA, because the pre-mRNA contains introns that are not needed for protein

synthesis, the introns are removed and the mRNA is formed containing only exons, through a process called splicing (Figure 2.3).

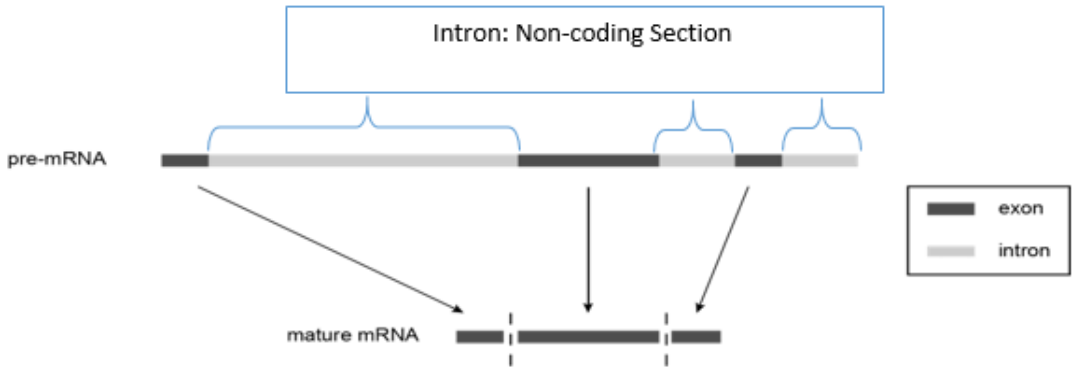


Figure 2.3: Splicing of introns from the pre-messenger RNA keeping only exons needed for translation (Figure adapted from [6]).

Transcription stages can also be divided into three stages namely; initiation, elongation and termination. The initiation stage begins with the binding of RNA polymerase to the DNA at the promoter at the beginning of a gene (Figure 2.4) (the sequence of promoter is as many as seven in eukaryotes but just three in bacteria). At the elongation stage, one of the strands of the DNA is taken as the template by the RNA polymerase, to make a new, complementary RNA molecule. RNA polymerase adds nucleotides to the 3' end. RNA is then synthesized in the 5' to 3' direction as in DNA replication. The process continues as the RNA polymerase advances until it reaches a certain sequence of nucleotides called the **terminator**. So as promoter indicates the start of transcription, terminator signals the end of it. At this stage, transcription stops as the new mRNA transcript and mRNA polymerase are released from DNA. As transcription is in progress, the DNA that has been transcribed re-winds to form a double helix (Figure 2.4).

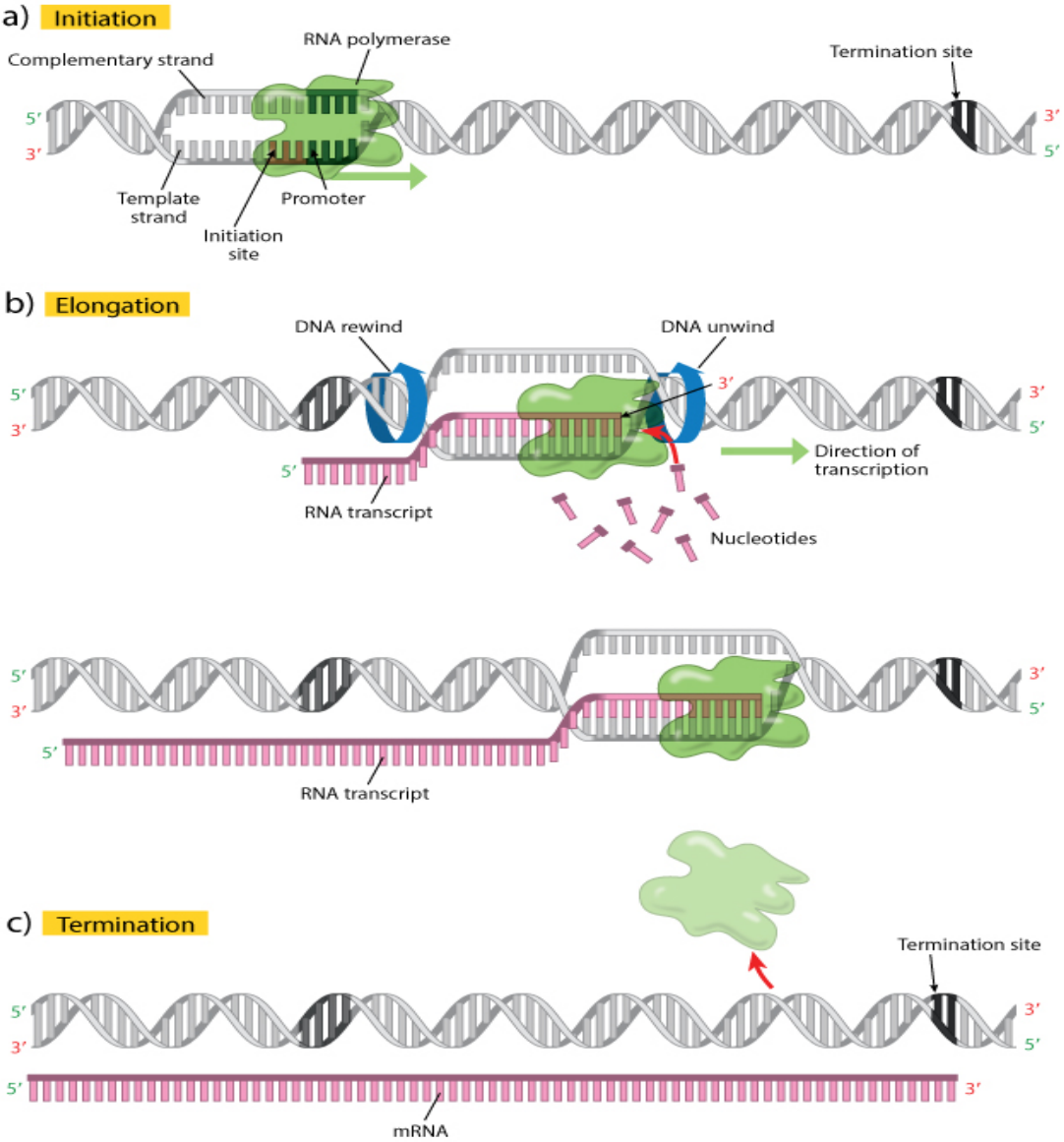


Figure 2.4: Transcription stages a) initiation, b) elongation c termination (Figure taken from [5])

2.1.2 Translation

Translation is the process by which protein is synthesized from the molecules of mRNA which have earlier been transcribed from DNA. To translate encoded mRNA into protein, mRNA has to be in the cytoplasm, where ribosome will aid translation. Ribosome is a huge complex of protein molecules and RNA. It is the site for translation and also the factory for protein synthesis. Transfer RNA (tRNA) is also needed for protein synthesis to assist mRNA in triggering protein synthesis. Protein

synthesis takes place in 4 phases; namely, initiation, elongation, termination and ribosome recycling. The first three phases are shown in Figure 2.5.

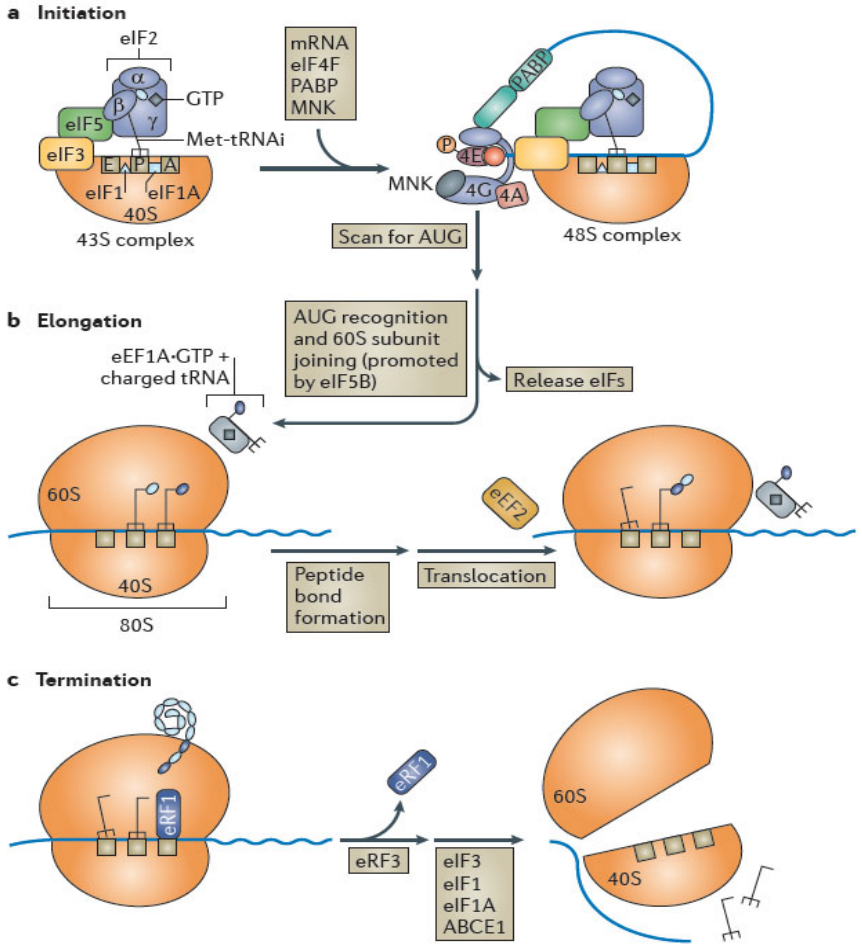


Figure 2.5: The first three phases of translation process a) initiation b) elongation c) termination (Figure taken from [7])

The initiation phase is the most complicated phase because it needs the highest number of protein factors compared to other phases. mRNA is triggered to move to the 40S ribosome at this stage, the start codon is located while the 60S ribosome attaches to the 40S ribosome to produce 80S ribosome, which is the elongation-complement.

The initiation stage is also the most regulated and is the mostrate-limiting step. The rate of limitation can also be different based on orders of magnitude which can stem from variation in mRNA regulatory features like untranslated region, highly structured 5´ or initiation regulation. This stage can be further divided into 5 steps, first is mRNA binding by the eukaryotic initiation factor 4F (eIF4F) cap-binding complex, which prepares the mRNA for translation. Second is 43S preinitiation complex (PIC) formation, and third is mRNA recruitment to the ribosome. The fourth step is initiation codon localization, while the last stage is the 60S ribosome attachment.

During elongation phase, the 80S ribosome moves on along mRNA, consequently translating into amino acid, all nucleotide triplet or codon. This codon is then fused with a developing polypeptide chain. Termination occurs when codon recognition stops.

Lastly, ribosome recycling takes place by releasing the mRNA while the 80S ribosome is split back into its original components of 40S and 60S. These can go on to be further recycled to start another process of translation [7].

2.1.3 Regulation of Gene Expression

As stated earlier, eukaryotic gene expression is the combination of the processes of transcription and translation. Gene expression is regulated at each of these levels as shown in Figure 2.6. At the transcription level, what gets transcribed can be regulated to get the primary transcript; then the number of exons versus the number of introns can be controlled. After splicing, what is exported from the nucleus can be regulated, how the exons are translated can further be regulated and finally when the

protein is made, it can be further modified, which can consequently change its shape and therefore its function [8].

At the transcription level, activities of polymerase which binds to DNA to initiate the process of transcription can be controlled in three main areas; Firstly, access to the gene is controlled; where polymerase access to the gene is controlled, which may include activities of enzymes that remodel histone. DNAs coil around this structure called histone, its modification can cause some part of the genome inaccessible to polymerases or their cofactors [9]. The rearrangement of histone to make it more accessible to polymerases and transcript factors is major transcription regulation process [10]. Secondly, elongation of the RNA transcript is regulated; that is the regulation of factors that allows the escape of polymerase from the promoter complex to begin transcribing RNA. Thirdly, regulation of the termination of polymerase, control of factors that determines when and how transcription termination occurs [11].

At translation level, most regulation occurs at the initiation stage. At this stage, under starvation or stress conditions, activation of mRNA for PIC binding by eukaryotic initiation factors (eIFs) can be controlled by inactivating these eIFs to reduce translation for most mRNAs. Translation initiation can also be blocked by reducing the activities of the eIFs that stimulate tRNA recruitment to the 40S subunit, which is essential for protein synthesis [12].

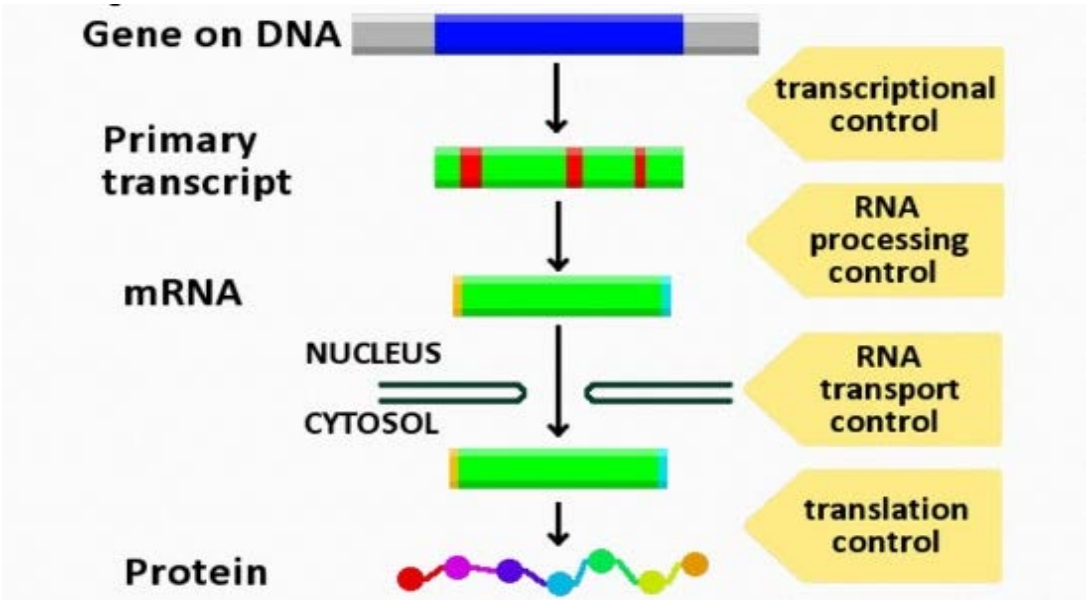


Figure 2.6: Levels of regulation of Gene Expression. (Figure taken from [13])

However, Gene expression is controlled both by extrinsic and by intrinsic factors. Intrinsic factors could include those mentioned in Figure 2.6. For example, a chromatin; which is the combination of a DNA and its associated histone proteins, can be altered chemically by a cell's own internal mechanism to change the ability of genes to access transcription factors, either positively or negatively. However, these changes do not modify the primary DNA sequence, this is to ensure that their daughter cells are compose of the same principal data at cell division.

Cell-extrinsic factors include environmental cues which could originate from either the organism's environment as well as from other cells of the organism because cells interacts with one another by sending and receiving growth factors (secreted proteins), and other signaling molecules. Exchange of signaling molecules between cells could cause semi-permanent changes in expression of genes. These changes in gene expression may be turning genes completely on or off, or cause a little reduction in the amount of transcript produced. Extrinsic factors could include small molecules, secreted proteins, temperature, and oxygen among others [14].

2.2 Overview of Bioinformatics

Bioinformatics could simply be defined as a methodology for biological analysis using computational techniques and algorithms with the aim of simplifying data representation with the aid of graphical and tabular representation. Bioinformatics deals with the collection, distribution and management of biological data. It combines different fields including, statistics, computer science, engineering, and mathematics, to analyze and to give simple interpretation to biological data. The techniques include data retrieval from various biological databases, analysis of the data retrieved and further processing with the aid of various software and algorithms [15].

The term bioinformatics was first used by Paulien Hogeweg and Ben Hesper in 1970 to mean “the study of information processes in biotic systems” [16], but the actual first step in Bioinformatics as it is known today, was the determination of sequence of insulins by Frederick Sanger in 1955 [17]. Table 2.1 shows in chronological order, a brief history of bioinformatics; including major development and innovations that have added to this area of science. These includes biological discoveries such as analyzing the first protein; innovations for sequencing and comparison in bioinformatics, such as BLAST and Entrez; as well as the establishment of biological databases such as NCBI [18] and PRINT protein database[19]. Apart from those included in the table, there has been significant growth in the field of bioinformatics, among which are those used in this work. Table 2.1 also includes information relevant to human, mouse and rat genomes, as they are analyses in thesis.

Table 2.1: A Chronological History of Bioinformatics (adapted from [16], [17], [20] and [21])

Year	Development in Bioinformatics	Developer(s)
1955	The sequence of the first protein is analyzed	Frederick Sanger
1970	Algorithm for sequence comparison is published.	Needleman-Wunsch
	The term Bioinformatics was coined	Paulien Hogeweg and Ben Hesper
1972	The first recombinant DNA molecule is created	Paul Berg
1973	The Brookhaven Protein DataBank is announced	
1985	The SWISS-PROT database is created	Department of Medical Biochemistry, University of Geneva and EMBL
1988	The National Centre for Biotechnology Information (NCBI) is established	
	The FASTA algorithm for sequence comparison is published	Pearson and Lupman

1990	The BLAST program is implemented	Michael Levitt and Chris Lee
	The human genome project started	
1994	The PRINTS database of protein motifs is published	Attwood and Bec
1999	Project to sequence the mouse genome is launched.	Mouse Genome Sequencing Consortium (MGSC)
2001	First drafts of the human genome are published	International Human Genome Sequencing Consortium
2002	The draft genome sequence for mouse is published.	
2003	Human Genome Project Completion	
2004	The draft genome sequence of Norway rat, Rattus norvegicus is completed	International Human Genome Sequencing Consortium

There are three important sub-disciplines of bioinformatics:

i. the development of new algorithms and statistics to check the relationships that exist among different data in a large data sets;

ii. the analysis and interpretation of varieties of data including protein domains, protein structures, nucleotide and amino acid sequences;

iii. and the development and implementation of tools that facilitates easy and efficient access and management of different types of data [22].

There are also three levels of bioinformatics

i. Single gene or protein analysis. This could include analyzing the sequence of a gene for similarity to other genes, features in the sequence, and prediction of secondary and tertiary structure.

ii. Genomes analysis: An entire genome is picked for analysis, which could be a check for which families of genes are present, location of genes in the chromosome as well their functions, and identification of missing enzymes in the genome.

iii. Analysis of genes and genomes with respect to functional data; such as analysis of a biochemical pathway and the identification of genes involved in an internal mechanism of an organism [22].

2.2.1 Biological Databases

There are basically two types of biological databases:

- **Archival (Primary) databases:** This may contain nucleic acid and protein sequences along with their annotations; compilation of mutations associated with diseases; organism based databases, such as specific genomes; databases focused on protein expression, metabolic pathways, regulatory networks and interactions. Examples of this type of databases include NCBI, and Ensembl; which are used in this work.

- **Derived (Secondary) databases:** These are made up of information retrieved as a result of analyzing archival databases. They may include sequence motifs such as characteristic patterns of families of proteins; classifications or relationships of features of entries in the databases; bibliographic databases such as PubMed [23]; databases of websites such as links between databases. These include Pfam [24], PROSITE [25] and PRINTS. [26].

2.2.2 Information Flow in Bioinformatics

This flow begins when scientists record and save results of experiments in a database. The data are then curated and annotated to ensure that they are properly stored in proper format for easy access in the future. Data is retrieved from the databases and analyzed for specific area of interest; discoveries are then published and stored in a database for future use.

Figure 2.7 shows an example of the progression of data/information in bioinformatics. As explained above, the results of a biological experiment such as, a protein sequence is stored in a biological database; this data is annotated in the database and made ready for public access. The saved data can be accessed later by an interested scientist; he/she may extract the relevant subsets of the data and then carry out analysis based on the areas of interest. The result of such analysis and/or experiments are aggregated according to homology, function and structure and then stored in a biological database, where they will be annotated and used further [27].

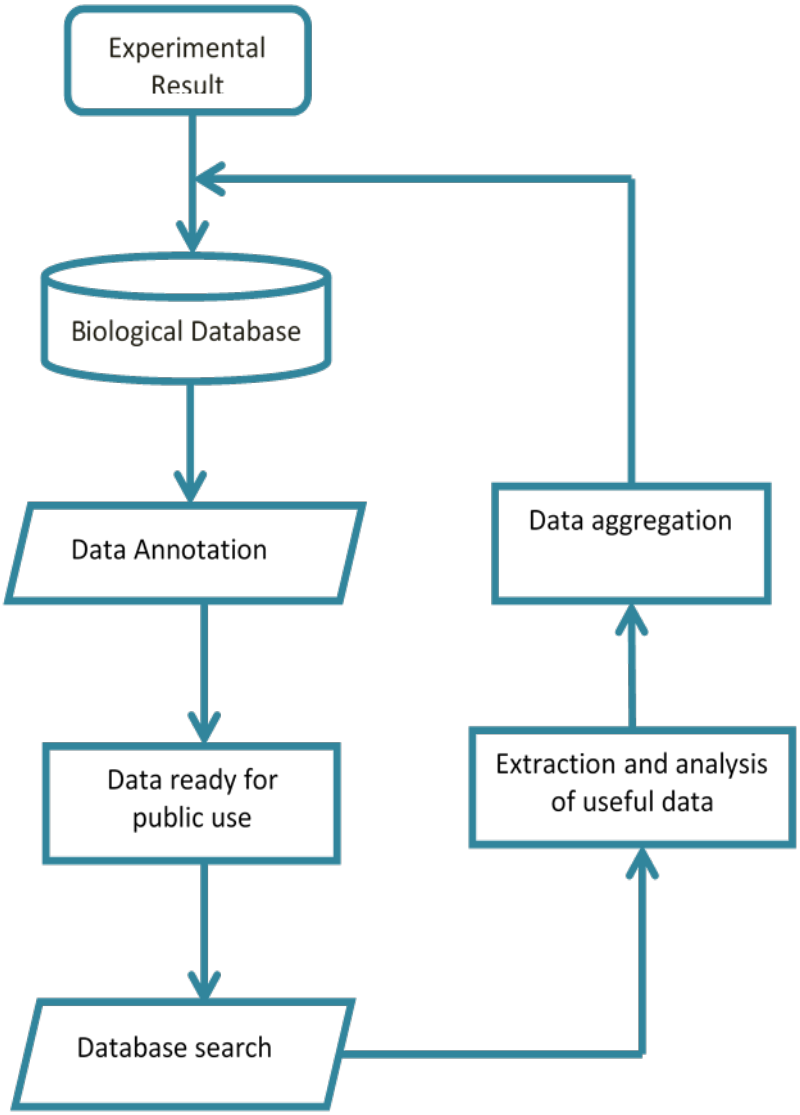


Figure 2.7: Basic Workflow for Bioinformatics Tasks

Chapter 3

G-PROTEIN-COUPLED RECEPTOR

G-protein-coupled receptors (GPCRs) sometimes called seven-transmembrane receptors (7TM) represent a group of diverse membrane receptors, which forms the largest and most diverse group of membrane receptors in eukaryotes. These receptors work as repository for messages which could be in the form of light energy, peptides, lipids, sugars, and proteins. The messages notify the cells of the availability of life-sustaining light or nutrients, or lack of these in their environment, they could also convey information received from other cells. Many eukaryotes depend on GPCRs to get information from their environment [26]. About 1000 GPCRs with specific signals are present in human. Understanding GPCRs is therefore important to modern medicine, because according to researchers, about one-out-of-three to one-out-of-two of drugs act by merging to GPCRs; this could increase because there are GCPRs whose ligands and physiological functions are not known, referred to as “orphan receptors”. Once they are “deorphanized”, a good number of them could be drug targets as well [28].

GPCRs interact with G proteins (proteins with the special ability to attach the nucleotides guanosine triphosphate (GTP) together with guanosine diphosphate (GDP)) in the plasma membrane. This is initiated when an external signaling molecule merges to a GPCR which leads to changes in the GPCR. G proteins which bind to GPCRs have 3 subunits, alpha, beta and gamma units. They are therefore referred to as **heterotrimeric** [28].

GDP binds to the alpha subunit whenever there is no signal, while the entire G protein-GDP complex attaches to a GPCR nearby until a signaling molecule gets to the GPCR. The signaling molecule causes a modulation of the configuration of the GPCR and consequently activates the G proteins while GTP takes the place of the GDP attached to alpha subunit as shown in Figure 3.1. These result in the dissociation of G protein subunit into 2 parts which are the GTP-bound alpha subunit and the beta-gamma dimer.

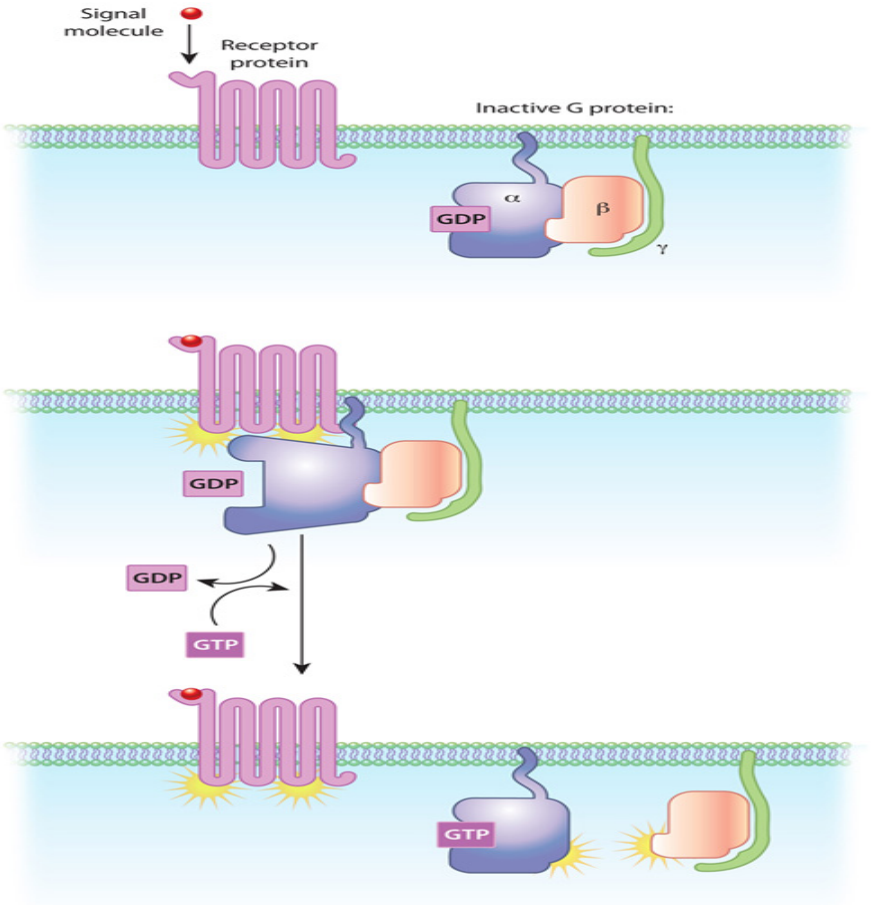


Figure 3.1: G-protein-coupled receptor activation process initiated by a signaling molecule (Figure taken from [28])

However, at this point, they are no longer attached to the GPCR although remain in the plasma membrane, where they are free to move and communicate with other

membrane proteins. While alpha subunits are attached to GTP, G proteins will stay active, at this time, both alpha subunit and beta-gamma dimer can interact with other membrane proteins to convert messages or energy to another in a cell.

G proteins are either excitatory (they trigger the activities of their target) or inhibitory (help to stop activities of such targets). G protein targets include enzymes which produce second messengers and ion channels, which give ions the ability to work as second messengers. Second messengers are tiny molecules that kick-start and monitors each intracellular signaling pathways. Examples of second messengers are cyclic AMP (cAMP) and diacylglycerol (DAG). cAMP is involved in many activities in the body such as responses to hormones, sensory input and nerve transmission. It is produced when an active G protein hits a target; adenylyl cyclase and activated by GTP-bound alpha subunit [28].

It is therefore clear that GPCRs is involved in a lot of internal mechanisms of organisms, ranging from sensation to hormone responses to growth, playing remarkable roles in sensing different signals from visual to olfactory. They help to establish sensory and regulatory connection between cell and external bodies, acting as receptors for outside ligands and as actuators for internal processes, thus, making the GPCR superfamily a major target for therapeutic intervention.

3.1 GPCR Groups

Classification of GPCR based on their amino acid sequences is very important due to the need to close the gap between large number of orphan receptors and the relatively small number of annotated receptors and because of research interest in these

sequences due to the importance of GPCR to modern drug industry and many other areas.

GPCRs can be grouped in five (5) major families

1. Class A (Rhodopsin family)
2. Class B (Secreting family)
3. Class B (Adhesion family)
4. Class C (Glutamate family)
5. Frizzled/TAS2 Family

Figure 3.2 shows all GPCR groups; Rhodopsin with the largest members, 701. Followed by Adhesion and Frizzled, with 24 each, and Secretin and Glutamate, with 15 each. Areas of close homologs of crystal structures with more than 35% sequence identity in the TM helices are highlighted in the figure. These areas are likely to be amenable for accurate comparative modeling.

The families are further divided into numerous subfamilies based on their sequence and sub-groups. Common subfamilies of Class A (Rhodopsin family) are shown in Figure 3.3. Rhodopsin family is classified into 19 subgroups/families while there are few unclassified GPCRs in this family. Other families equally have different subfamilies as in the case of Class A [29].

The families have very little sequence similarity (SS) of less than twenty percent (SS, < 20% in the transmembrane (TM) domain) and their extracellular N-terminal domains are different. For example, Class A consists of about 700 GPCRs in humans, which are further classified into four subgroups (these subgroups have SS of

more than twenty five percent (SS ≥ 25%)). Each subgroup also have numerous subfamilies that share higher sequence similarity of more than 30% (SS ≥ 30%) [30].

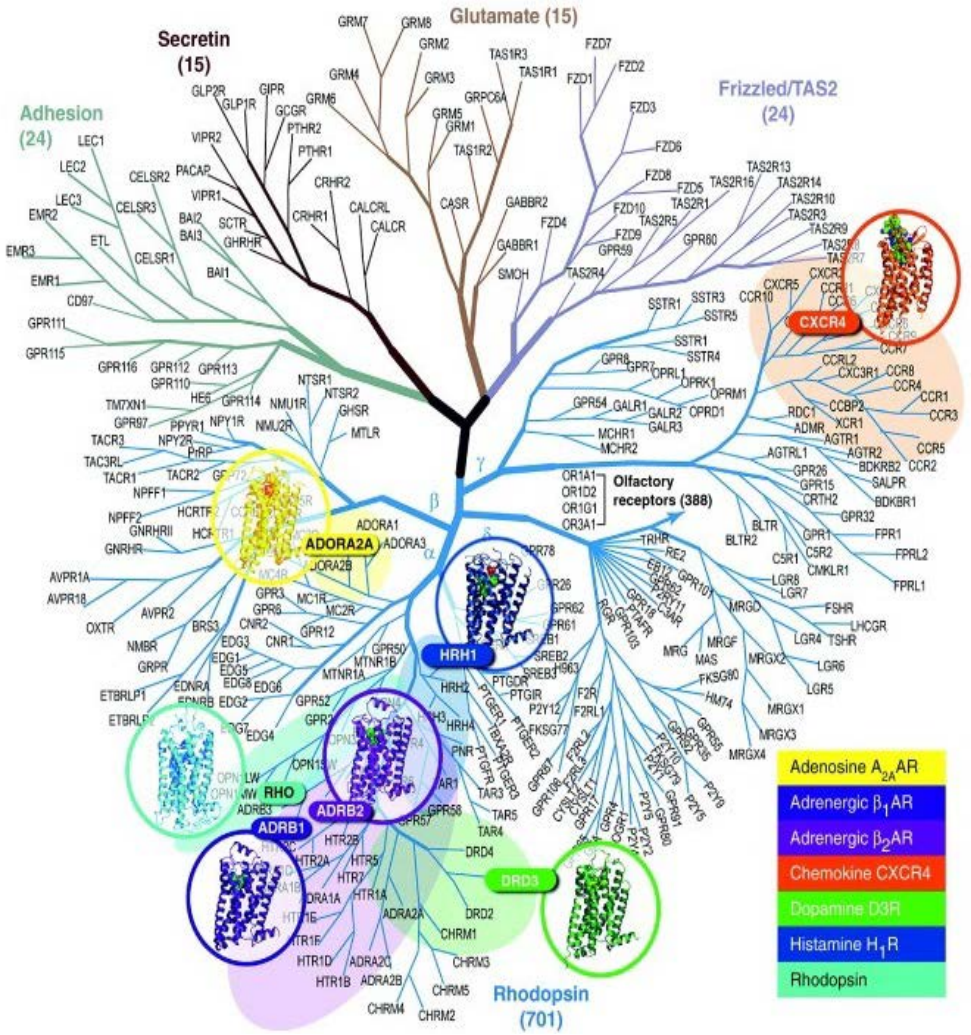


Figure 3.2: GPCR family tree showing all 5 major families (Figure taken from [31])

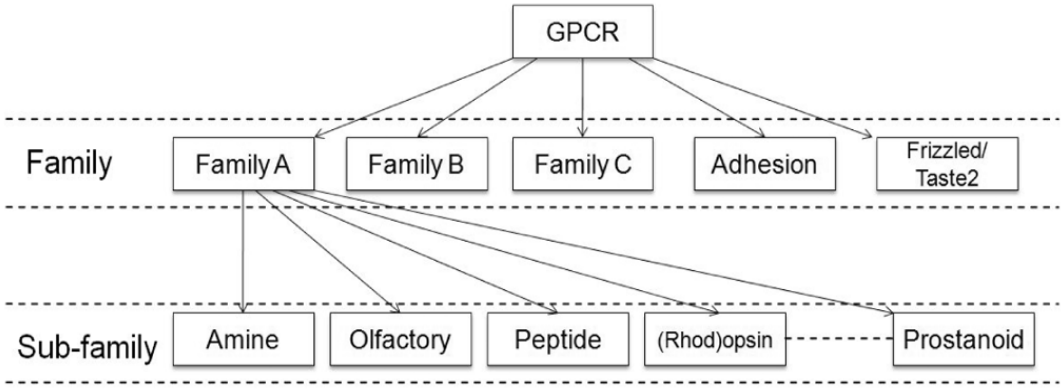


Figure 3.3: GPCR families and sub-families of Rhodopsin (Figure taken from [29])

3.2 GPCR Structure

Research on the structures of GPCRs has received a dramatic boost in recent time with breakthroughs in GPCR crystallography, giving hope that structural mysteries of majority of subfamilies will be solved in the next few years [31].

All GPCRs have a common seven transmembrane (7TM) topology; however, there is great variety of features, dynamics, selectivity to ligands, modulators and downstream signaling effectors in their structure. The greatest structural differences can be found among GPCR classes and subfamilies, but structural and sequence similarity are high enough within classes and subfamilies to allow for accurate predictions by comparative modeling of protein, (that is the construction of an atomic-resolution model a "target" protein from its amino acid sequence and an experimental three-dimensional structure of a related homologous protein). This is used in applications such as ligand docking, virtual screening for dopamine D3 antagonists, and profiling of ligand selectivity within the adenosine receptor subfamily [31].

The 7TM bundle of GPCRs is connected by three extracellular loops (ECL); responsible for ligand binding and three intracellular loops (ICL); responsible for downstream signaling, interacting with G proteins and other effectors in the same region, as shown in Figure 3.4. The extracellular (EC) part include N-terminus which ranges from often unstructured and short sequences in Class A to large globular EC part in other classes of GPCR. The intracellular (IC) part usually includes helix VIII

and a C-terminus sequence that often carrier of signal sites such as Palmitoylation, which is the covalent attachment of fatty acids to cysteine and to serine and threonine residues of proteins (though less frequently), which are typically membrane proteins [29].

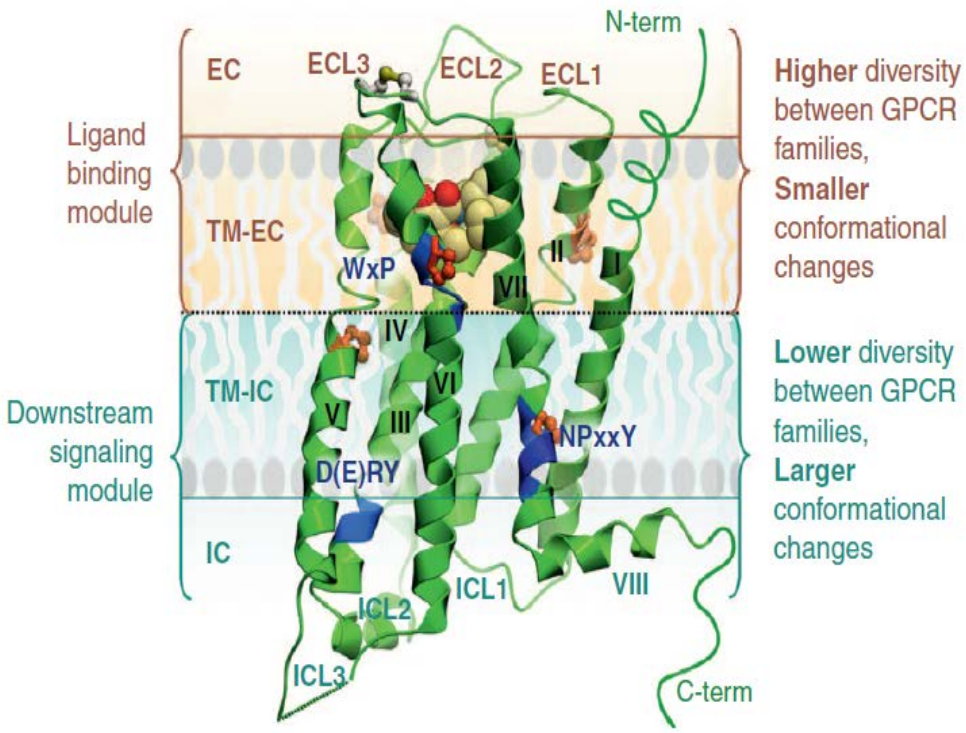


Figure 3.4: General structure of GPCRs comprising extracellular (EC) and intracellular (IC) parts (Figure taken from [31])

The 7TM helical bundle which is recognized as the most conserved component of GPCRs shows characteristic hydrophobic patterns and houses signature motifs that are functionally important such as the D[E]RY motif in helix III (part of the so-called ‘ionic lock’), the WxP motif in helix VI, and the NPxxY motif in helix VII. Crystal structure of Class A GPCRs show the overall structural conservation of 7TM fold to be true while also revealing obvious structural diversity in both the loop regions and the helical bundle itself. Although the variations are more pronounced on the EC than on the IC side of the receptors crystallography results show that the most

important variations are in the extracellular loop region, where stock of secondary structure or types and disulfide crosslinking are presented. The 7TM helical bundle itself also has important variations.

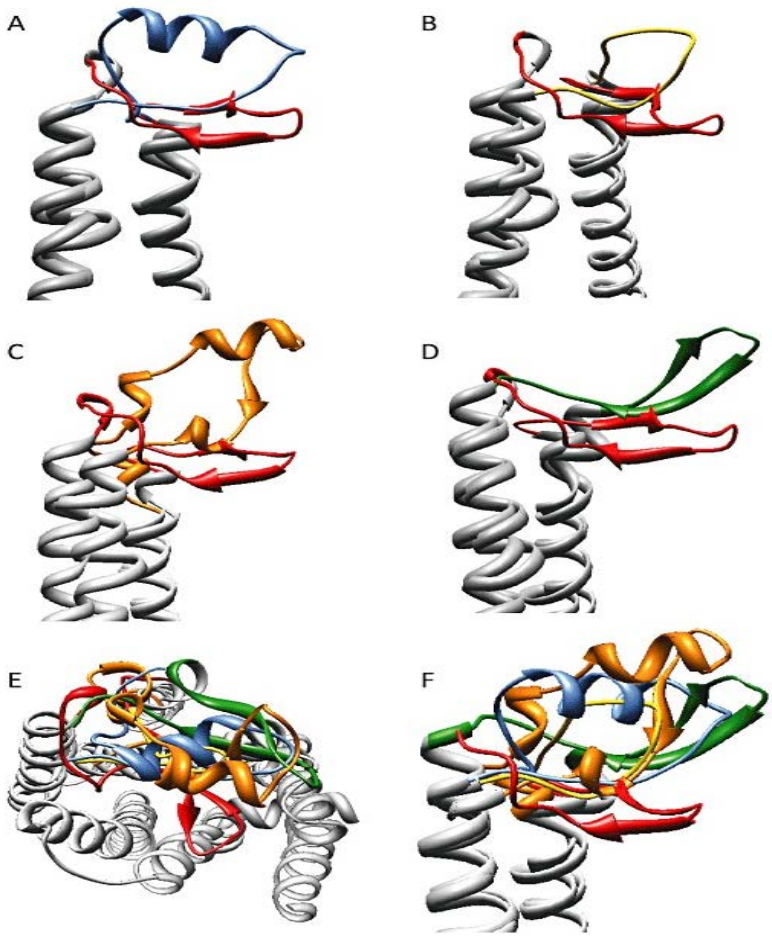


Figure 3.5: Differences in ECL2 region of GPCR (Figure taken from [32])

Figure 3.5 shows the diversity in the ECL2 of GPCRs. The ECL2 region is usually the longest of the ECL, though not always, and it is where most of the diversity is observed. Rhodopsin (shown in red color) is compared to four different diffusible-ligand GPCRs. Panel A in the figure shows adrenergic receptors ($\beta_{2A}R$) compared to rhodopsin; panel B compares dopamine receptor (D3R) with it; panel C compares adenosine receptor ($A_{2A}R$)with it; while panel D compares chemokine receptor type

4 (CXCR4) with it. Panel E shows an overlay of ECL2 of all 5 GPCRs viewed from above while F shows the view from the plane of the membrane.

ECL2 in rhodopsin is made up of two β-sheets (β3 and β4) which interact with β1 and β2 in its structured N-terminal region. They form a β-hairpin that plunges downward onto the TM bundle as shown in Figure 3.5A. On the other hand, β_{2A}R has unstructured N-terminal region and its structure is a short α-helix structure that is stabilized by an intra-helical bond. Other GPCRs shown equally has different structures [32].

ECL2 and the N-terminus in rhodopsin forms a lid over the binding pocket protecting the pre-bound ligand, but in β_{2A}R, D3R, A_{2A}R and CXCR4, ECL2 lies more peripheral to the binding crevice entrance as shown Figure 3.5E, F. From the figure, it is can be concluded that ECL2 conformation is different across GPCRs. These structural differences translate into functional differences; initially originated by transcript and protein domain diversity.

Chapter 4

METHODOLOGY

The goal of this thesis is to analyze the impact of transcript diversity on protein domains of G-protein-coupled Receptors (GPCRs) in three different genomes (*Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*). This was done by first searching databases for relevant data, and retrieving information related to the topic from biological databases such as NCBI and Ensembl. The data collected was stored in a newly constructed database and finally analyzed using bioinformatics tools. There has been a surge in the number of biological databases and tools used to retrieve and analyze data. The following sections will discuss the databases and the tools used to analyze the data in this thesis.

4.1 Biological Databases and Resources Used

A number of public biological databases and resources were used in this work, including NCBI, Ensembl [2], BioMart [33], Pfam and Uniprot [34]. They contain different form of data/information that are relevant to this work and bioinformatics in general. These databases and resources used are described in the following sections.

4.1.1 National Center for Biotechnology Information (NCBI)

NCBI was founded in 1988 to house databases related to biotechnology and biomedicine, which are very important for bioinformatics. Located in Bethesda, Maryland USA, NCBI is a department under the National Library of Medicine which

is a branch of National Institutes of Health of the United State. NCBI has been making DNA sequence database (GenBank) available to scientists since 1992 as well as coordinating with other similar databases such as the DNA Data Bank of Japan (DDBJ) and the European Molecular Biology Laboratory (EMBL) [35].

NCBI provides tools like BLAST and Entrez to make analysis of data in the GenBank easier for users all of which can be accessed from its homepage shown in Figure 4.1.

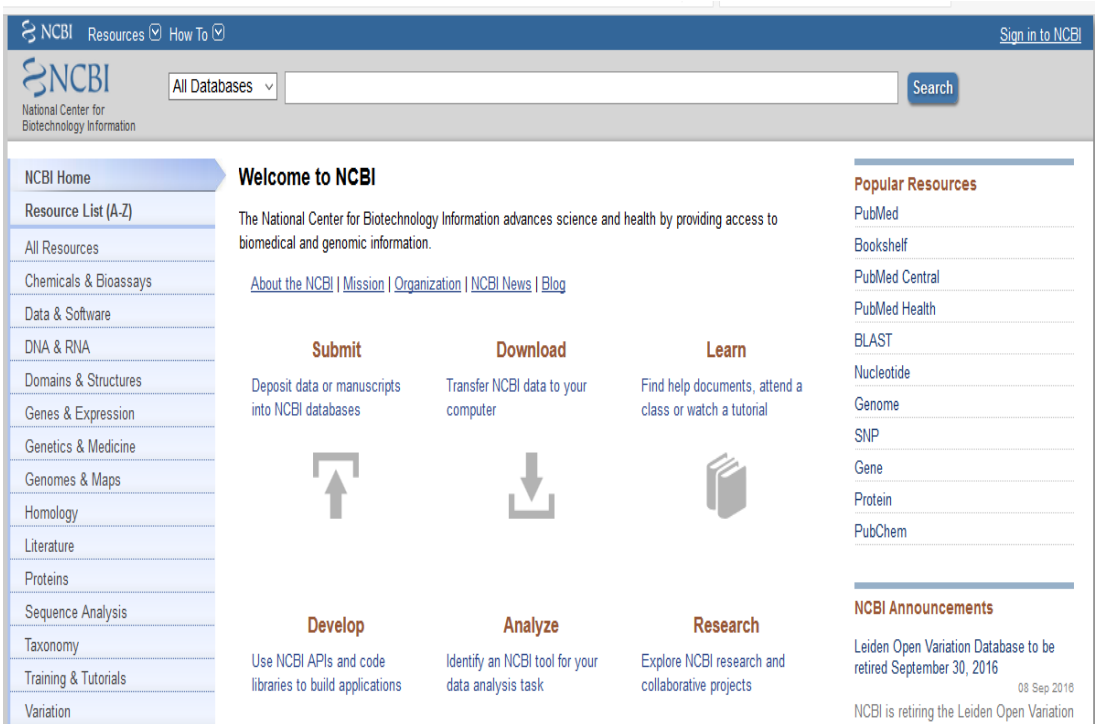


Figure 4.1: A snapshot from the NCBI homepage

BLAST (Basic Local Alignment Search Tool) is a search tool on the website of NCBI, with features designed to make searching for specific area of interest in the database easy. It is used to filter out results as required by the user. BLAST has a standalone application that can be downloaded and

installed on a PC with full features, hence, a complement of the website, for easy access and data analysis [36].

BLAST provides specialized searches such as SmartBLAST which finds all proteins similar to query entered, Primer-BLAST which designs primers according to a specified template. Global Align which compares two sequences across their entire span and the likes.

A new Application Programming Interface (API) called Magic BLAST is now being introduced as an improved tool for mapping large sets of next-generation RNA or DNA sequencing runs against a whole genome or transcriptome. It optimizes score of inputs, locates its introns and adds up the score for all exons using NCBI BLAST libraries. It also gives sequence results in FASTA, SRA files or NCBI SRA accession formats. Magic BLAST executables are available for LINUX, MacOSX, and Windows. The tool is under active development and new releases are expected from time to time [37].

Entrez: Entrez provides an alternative platform on NCBI where search engine forms can be used to query data. More importantly, Entrez provide Entrez Programming Utilities (E-utilities), a set of eight server-side programs which provide users with direct access to up to 38 databases to search and retrieve requested data using fixed URL syntax. This syntax can be used in different programming languages such as Perl to provide access to all databases simultaneously with a single query string [38].

The E-utilities include, EInfo (database statistics), ESearch (text searches), EPost (UID uploads), ESummary (document summary downloads), EFetch (data record downloads), ELink (Entrez links), EGQuery (global query), ESpell (spelling suggestions), ECitMatch (batch citation searching in PubMed) [39].

4.1.2 Ensembl

The Ensembl project was initiated in 1999 due to the major growth in the number of sequences that are being stored in databases. Since working with such large data would be an overwhelming task, Ensembl was launched in 2000 to annotate the genome, integrate this annotation with other available biological data automatically as well as make them available to the public through the website which is publicly available via the web http://www.ensembl.org. The homepage of Ensembl is shown in Figure 4.2. The human genome was the first to be available on this project, but many more have since been added which led to the creation of sister websites to serve specific genomes.

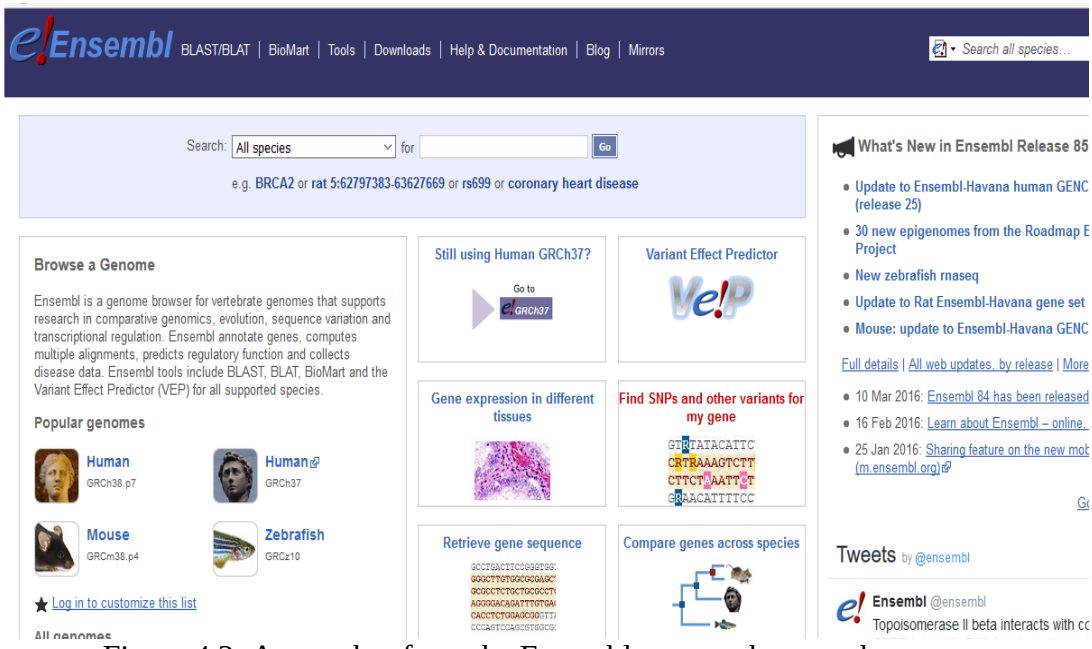


Figure 4.2: A snapshot from the Ensembl genome browser homepage

With over 1000 databases in biological fields, there is the need to develop tools to search through these databases and to process data. Ensembl provides ready-made tools for users to processes data on the databases as well as users' results. These tools are categorized into two, data processing tools and tools for accessing Ensembl data.

4.1.2.1 Data processing tools:

- **Variant Effect Predictor**: Analyse user's variants and predict the functional consequences of known and unknown variants.
- **BLAST/BLAT**: Search through genome databases on Ensembl for DNA or protein sequence inputted by the user.
- **File Chameleon**: Help to convert Ensembl files for use with other analysis tools which are usually standalone API.
- **Assembly Converter**: Used to map user's annotation files to the current assembly using CrossMap which is a program that converts genome coordinates between different assemblies (such as between Human genomes hg18 (NCBI36) and hg19 (GRCh37).
- **ID History Converter**: Convert Ensembl IDs of a previous release to their current equivalents.

4.1.2.2 Accessing Ensembl data tools:

- **Ensembl Perl API**: Uses Perl scripts to access all Ensembl data.
- **Ensembl Virtual Machine**: VirtualBox which is a virtual machine with Ubuntu desktop and pre-configured with the latest Ensembl API plus variant effect predictor for easy access to Ensembl databases without a need for a browser.
- **Ensembl REST server**: This gives users the opportunity to choose their own programming language with which they wish to access Ensembl databases.

- **BioMart**: This is used to export customized datasets from Ensembl. BioMart provides a platform to mine Ensembl databases conveniently according to the interest of the user. Figure 4.3 shows the BioMart page and short description of how it can be used to search for data and give the results in tabular form according to the interest of the user [40].

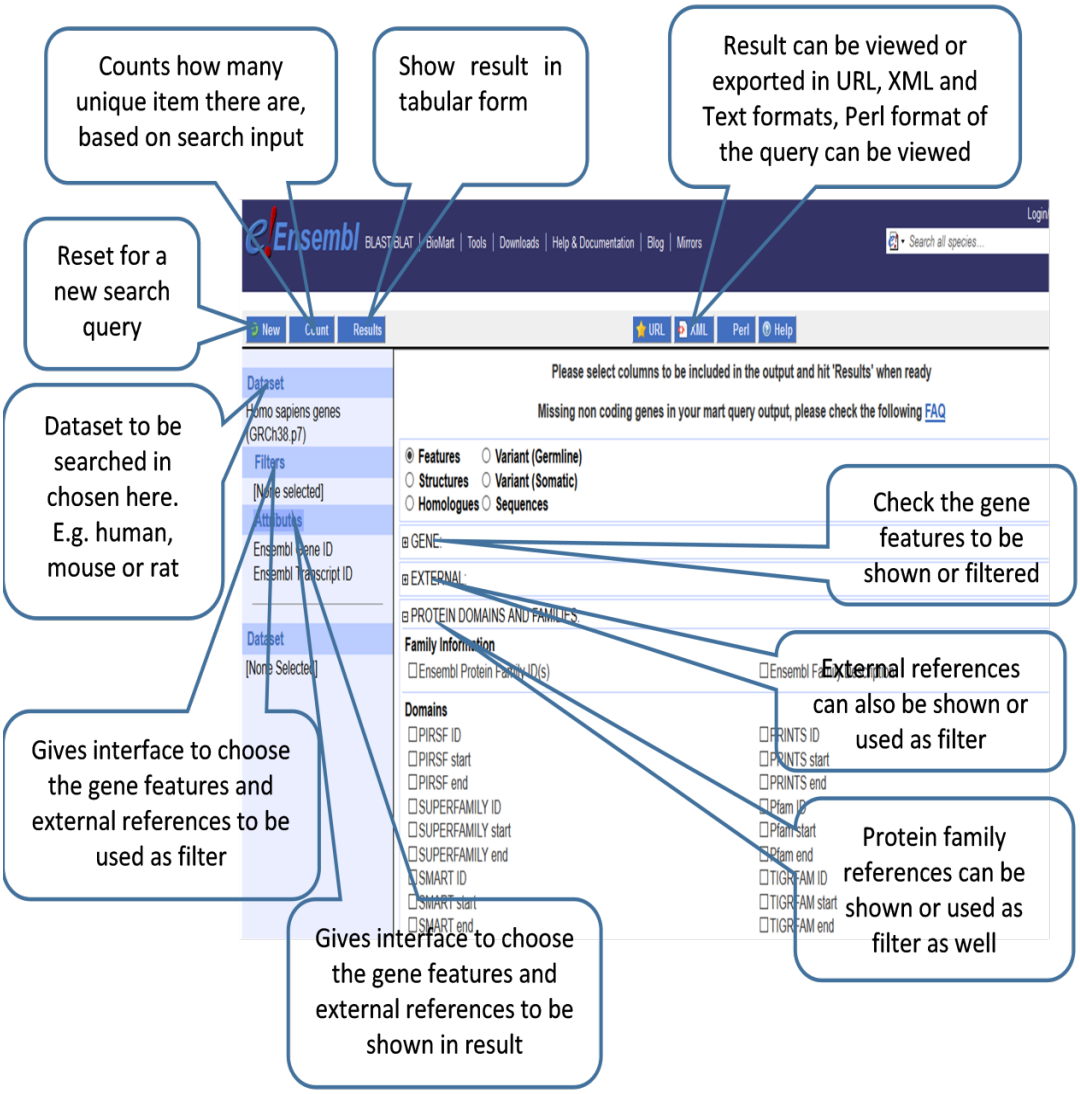


Figure 4.3: A sample BioMart interface

Users can choose from all available datasets, the genome of interest to them (such as *Anas platyrhynchos* genes, *Homo sapiens* genes, and *Mus musculus* genes). Attributes that are required to be present can be selected, ranging

from features, variants, structures, homologues to sequences. These attributes can be further chosen using “filters” such as specifying a region or a gene of interest, domain or domain diversity, phenotype or gene ontology [41].

4.1.3 Protein Family (Pfam)

Pfam is a sequence (Pfamseq) database of protein families which contain around 15,000 entries defined by profile Hidden Markov model (HMM). This is a model based on probability for statistical analysis of homology with the aim of producing protein families that successfully classify sequence spaces with high accuracy. Pfam, developed by European Bioinformatics Institute (EMBL-EBI) is available as a free online resource available on http://pfam.sanger.ac.uk/ or (http://pfam.janelia.org/. It provides domain graphics, which are graphical representations of search results using domain graphic generator. Figure 4.4 gives short descriptions of different functions and searches that can be carried out using the Pfam homepage.

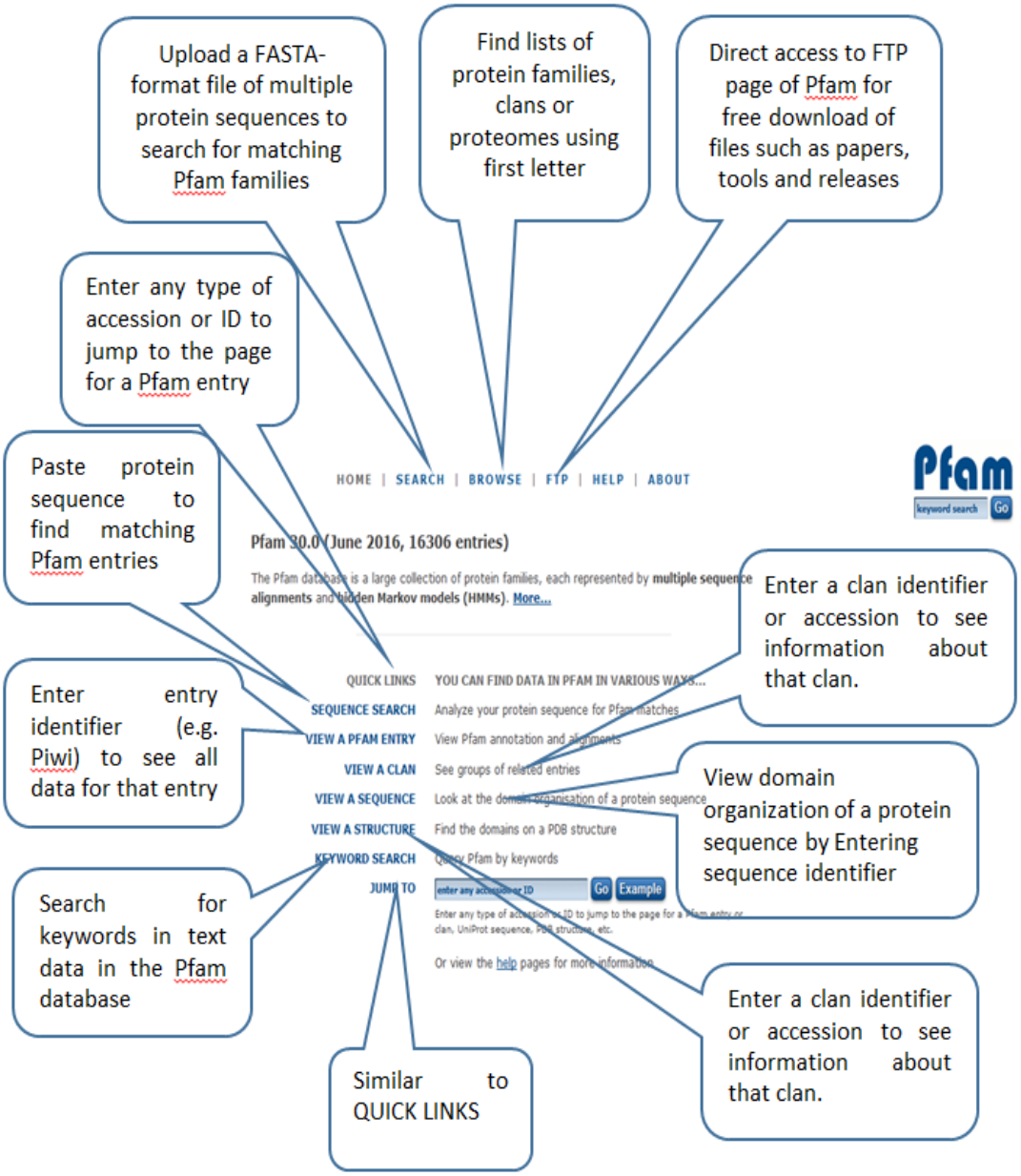


Figure 4.4: Pfam family web page

4.1.4 Universal Protein Resource (UniProt)

UniProt database is the collaboration between EMBL-EBI, Swiss Institute of Bioinformatics (SIB) and Protein Information Resource (PIR) with the main aim of providing databases which comprehensively cover protein sequence and annotation data. Similar to other biological databases, it is linked to other databases like Ensembl by UniProtKB identifier. Figure 3.5 shows a typical UniProt webpage where different interfaces can be used depending on the user's requirement.

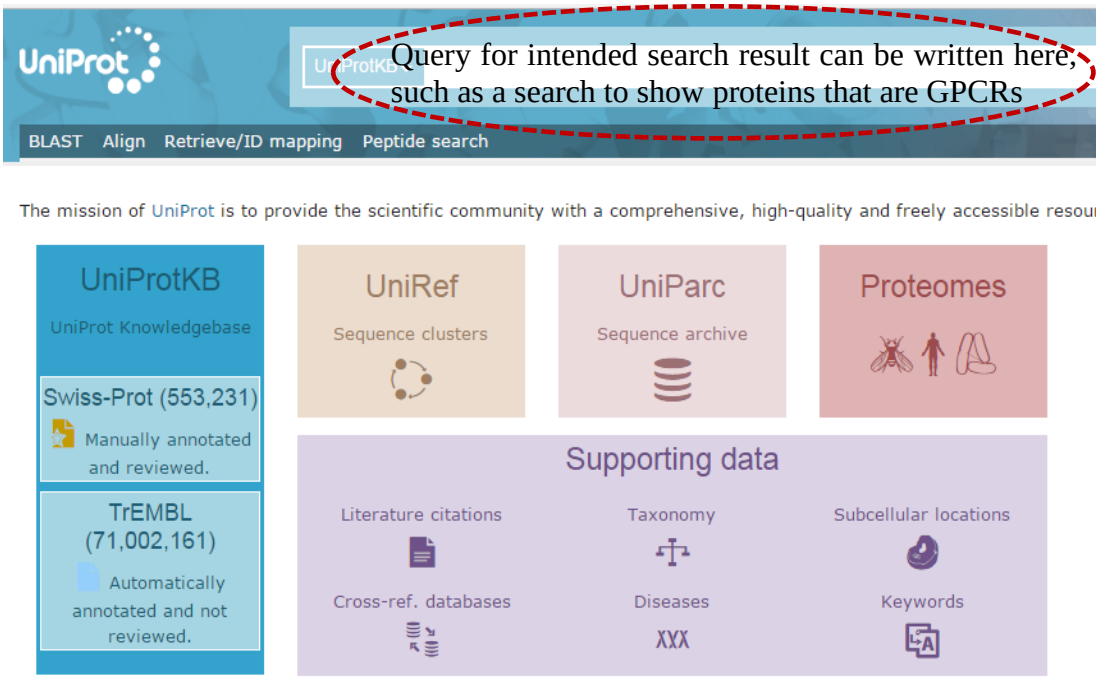


Figure 4.4: A typical UniProt webpage

The features on UniProt include:

- **BLAST:** Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences which can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.
- **Align:** Used to align two or more protein sequences with the Clustal Omega program (a multiple sequence alignment program for proteins which produces biologically meaningful multiple sequence alignments of divergent sequences) to view their characteristics alongside each other.
- **Retrieve/ID mapping:** List of identifiers can be entered or uploaded here either to retrieve the corresponding UniProt entries to download or to work with them on the website. It also helps to convert identifiers which are of different types to UniProt identifiers or vice versa and download the identifier lists.

- **Peptide search**: Search tool for finding all UniProtKB sequences that exactly match a query peptide sequence.

4.2 Tools used

Data retrieved from the biological databases must be saved in a private database for easy access at a later time. Also downloaded files from these databases often contain more information that required so there is a need to filter out information irrelevant for a particular task before storing the data in the constructed private database. For this work, we used PhpMyAdmin as the interface for saving our data and Perl (using **Strawberry Perl** as interface) is used as the programming language for parsing the XML file downloaded which contains data from biological databases.

- **PhpMyAdmin** got its name from its function as a tool which uses PHP language in MySQLdatabases to manage and administer the activities of users. It is a free and open source tool developed in 1998 by Tobias Ratschiller but has since been modified and approved with several releases based on the initial work of Tobias Ratschiller. This was aimed at providing an easy platform to create, modify or delete data from databases, tables and management of users and their corresponding permissions to the data or databases [42].

  XAMPP incorporates all the features of phpMyAdmin as well as other useful software. It is a free and open source software used as a server for local hosting on a system made possible by its light-weighted Apache server. It is a cross-platform web server which derives its name from its functions: X for Cross-Platform, A for Apache, M for MySQL, P for PHP and the last P for Perl. It is light-weighted Apache server makes it very easy for developers to create a local http server with just few clicks, as well as creating databases and other functions

[43] such as the one used in this thesis, where the functions of phpMyAdmin are widely employed for the exploration of data.

- **Strawberry Perl**: Perl language is generally designed to work on UNIX systems, but Strawberry provides an easier environment for Microsoft Windows users as it contains all the functions needed to run and develop Perl applications, thereby working as close as possible to Perl environment on UNIX systems [44].

4.3 Data Retrieval and Organization

The first step in this work is to search for relevant data in biological databases and download the file containing the data from the related website. This is done separately for each one of the three species (human, mouse and rat) analyzed in this work. The data is further "filtered" after downloading based on the requirements of this work. The final data are then stored in the database constructed separately for the three species.

4.3.1 Data Retrieval

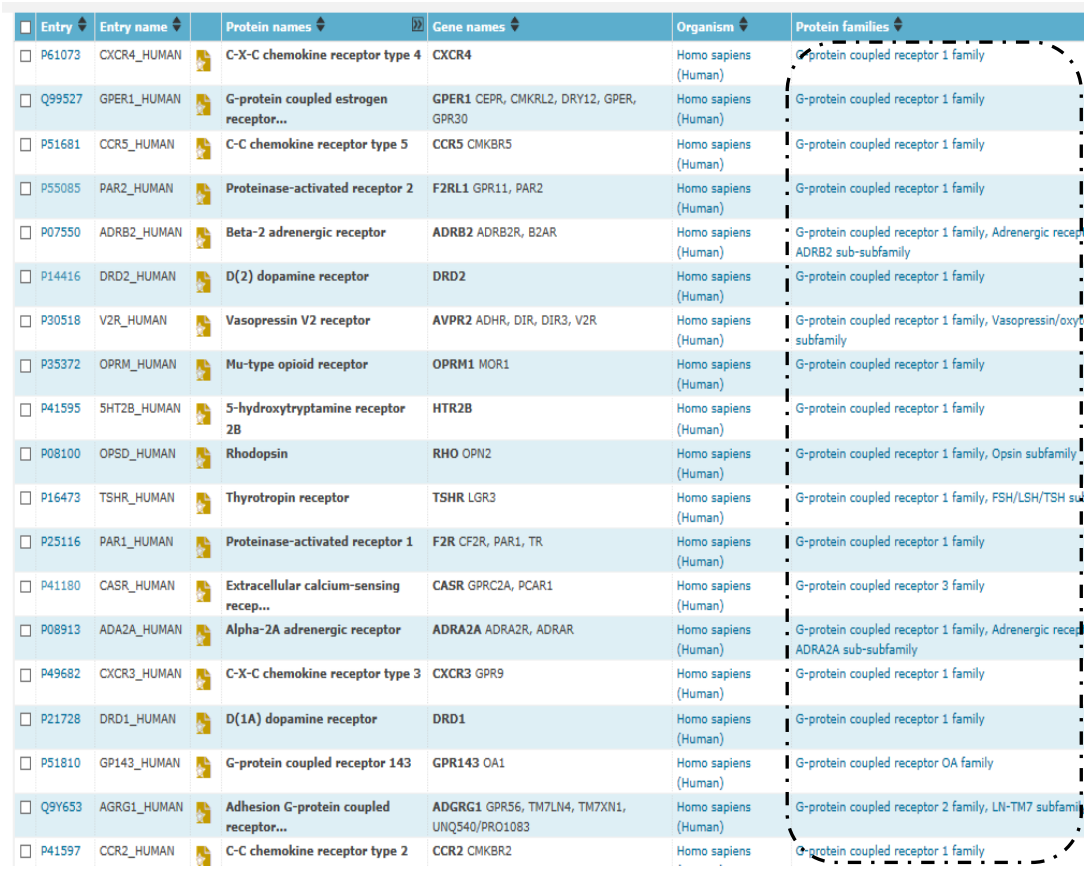
Gene data are stored in several popular databases in the domain such as NCBI and UniProt. Since data are being updated across different databases continuously, it is important to retrieve the most comprehensive and up to date data. For this purpose, the UniProt database is chosen after analyzing databases such as NCBI, Ensembl, Reactome, and GPCRdb and concluding that UniProt gives the most comprehensive result. Data search is done in UniProt for the three species using the queries given in Table 4.1.

Table 4.1: Queries and results in UniProt

Organism	**Query**
Human	*"family: 'G-protein coupled receptor' and organism: human and reviewed: yes"*
Mouse	*"family: 'G-protein coupled receptor' and organism: mouse and reviewed: yes"*
Rat	*"family: 'G-protein coupled receptor' and organism: rat and reviewed: yes"*

Figure 4.6 shows a screenshot of the result of querying UniProt for human species. The protein family column confirms that the results belong to G-protein coupled receptors as desired. However, our study focuses on genes rather than proteins. Therefore, the data obtained is used later to search Ensembl Biomart database, where result are obtained for the desired genes.

Result from UniProt is retrieved from the database in XML format for the three species. These files contain all the data related to each protein and protein families in our search category. However, only UniProtIDs are needed, which are used as filters in querying Ensembl Biomart database. A simple Perl code is written to parse the XML files, to keep only the UniProtIDs of the proteins in the files. The Perl code used for this is given in Appendix A.



Entry	Entry name	Protein names	Gene names	Organism	Protein families
P61073	CXCR4_HUMAN	**C-X-C chemokine receptor type 4**	**CXCR4**	Homo sapiens (Human)	G-protein coupled receptor 1 family
Q99527	GPER1_HUMAN	**G-protein coupled estrogen receptor...**	**GPER1** CEPR, CMKRL2, DRY12, GPER, GPR30	Homo sapiens (Human)	G-protein coupled receptor 1 family
P51681	CCR5_HUMAN	**C-C chemokine receptor type 5**	**CCR5** CMKBR5	Homo sapiens (Human)	G-protein coupled receptor 1 family
P55085	PAR2_HUMAN	**Proteinase-activated receptor 2**	**F2RL1** GPR11, PAR2	Homo sapiens (Human)	G-protein coupled receptor 1 family
P07550	ADRB2_HUMAN	**Beta-2 adrenergic receptor**	**ADRB2** ADRB2R, B2AR	Homo sapiens (Human)	G-protein coupled receptor 1 family, Adrenergic recep ADRB2 sub-subfamily
P14416	DRD2_HUMAN	**D(2) dopamine receptor**	**DRD2**	Homo sapiens (Human)	G-protein coupled receptor 1 family
P30518	V2R_HUMAN	**Vasopressin V2 receptor**	**AVPR2** ADHR, DIR, DIR3, V2R	Homo sapiens (Human)	G-protein coupled receptor 1 family, Vasopressin/oxyt subfamily
P35372	OPRM_HUMAN	**Mu-type opioid receptor**	**OPRM1** MOR1	Homo sapiens (Human)	G-protein coupled receptor 1 family
P41595	5HT2B_HUMAN	**5-hydroxytryptamine receptor 2B**	**HTR2B**	Homo sapiens (Human)	G-protein coupled receptor 1 family
P08100	OPSD_HUMAN	**Rhodopsin**	**RHO** OPN2	Homo sapiens (Human)	G-protein coupled receptor 1 family, Opsin subfamily
P16473	TSHR_HUMAN	**Thyrotropin receptor**	**TSHR** LGR3	Homo sapiens (Human)	G-protein coupled receptor 1 family, FSH/LSH/TSH su
P25116	PAR1_HUMAN	**Proteinase-activated receptor 1**	**F2R** CF2R, PAR1, TR	Homo sapiens (Human)	G-protein coupled receptor 1 family
P41180	CASR_HUMAN	**Extracellular calcium-sensing recep...**	**CASR** GPRC2A, PCAR1	Homo sapiens (Human)	G-protein coupled receptor 3 family
P08913	ADA2A_HUMAN	**Alpha-2A adrenergic receptor**	**ADRA2A** ADRA2R, ADRAR	Homo sapiens (Human)	G-protein coupled receptor 1 family, Adrenergic recep ADRA2A sub-subfamily
P49682	CXCR3_HUMAN	**C-X-C chemokine receptor type 3**	**CXCR3** GPR9	Homo sapiens (Human)	G-protein coupled receptor 1 family
P21728	DRD1_HUMAN	**D(1A) dopamine receptor**	**DRD1**	Homo sapiens (Human)	G-protein coupled receptor 1 family
P51810	GP143_HUMAN	**G-protein coupled receptor 143**	**GPR143** OA1	Homo sapiens (Human)	G-protein coupled receptor OA family
Q9Y653	AGRG1_HUMAN	**Adhesion G-protein coupled receptor...**	**ADGRG1** GPR56, TM7LN4, TM7XN1, UNQ540/PRO1083	Homo sapiens (Human)	G-protein coupled receptor 2 family, LN-TM7 subfami
P41597	CCR2_HUMAN	**C-C chemokine receptor type 2**	**CCR2** CMKBR2	Homo sapiens	G-protein coupled receptor 1 family

Figure 4.5: Screenshot of a result from UniProt (data retrieved in November, 2016)

Biomart search is done to generate a tabular representation of the needed genes. Attributes chosen for each gene to be represented contain Ensembl Gene ID, Ensembl Transcript ID, Pfam ID, transcript type and transcript count (i.e. the number of transcripts that a particular gene has). The filters applied include the following:

- Transcript count should be greater than 1 (only genes with multiple transcripts should be considered).
- Transcript type should be protein-coding.
- Only genes whose UniProtIDs were retrieved in the previous search are included.

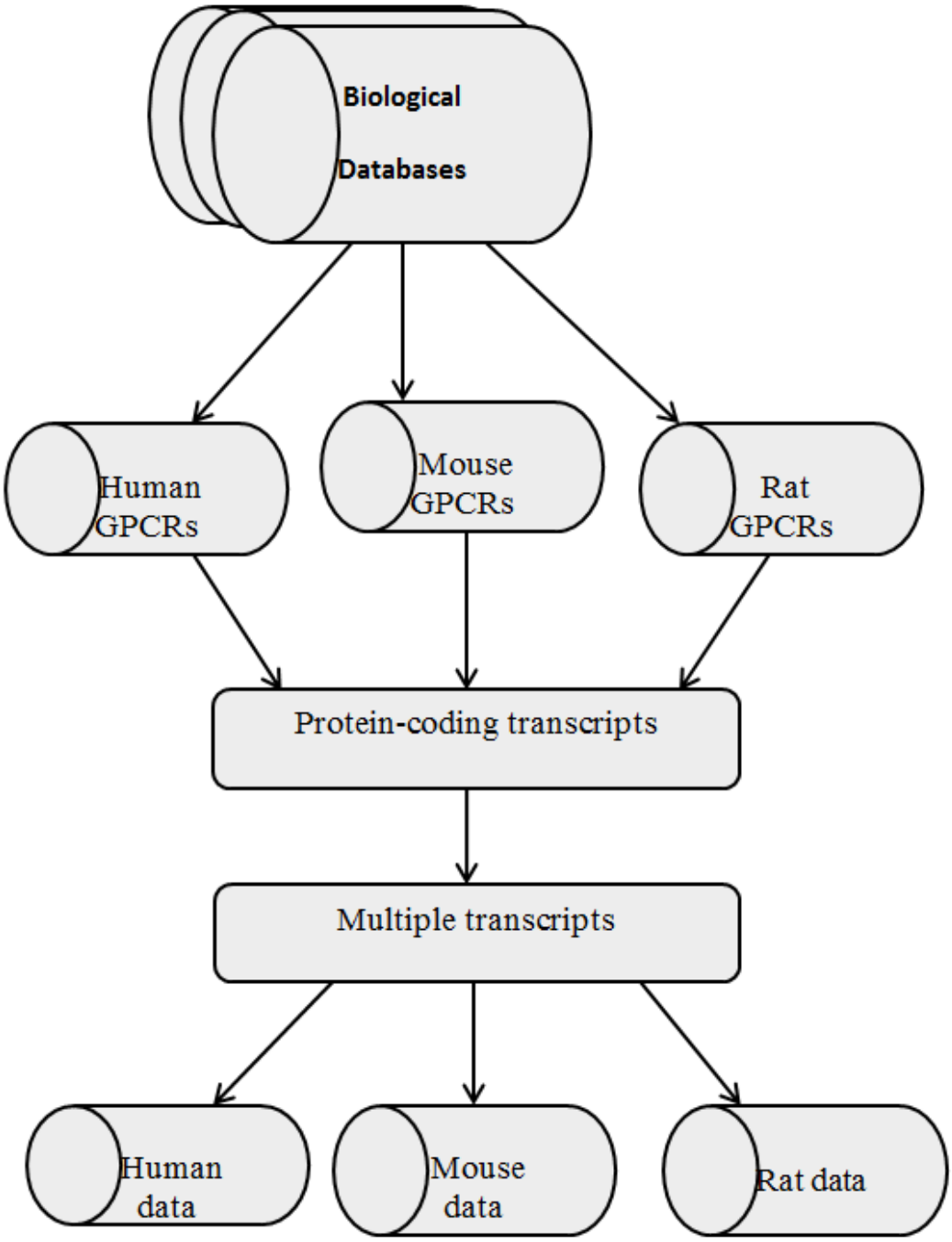


Figure 4.6: Data Retrieval Stages from Biological Databases for 3 Species

Results are retrieved and stored in a database created mainly for this work to be described in the next section. Figure 4.7 shows the stages used to collect the final data used in this work. First, a genome is chosen (i.e. human), and then searched for only GPCRs. After this, genes which are non-protein coding are excluded and finally, only those with multiple transcripts are stored.

4.3.2 Database Constructed

It is important to store the retrieved data in a relational database in order to efficiently process the data and generate results. Therefore, a relational database is designed and created as part of this thesis. PhpMyAdmin incorporated in XAMPP server was used for this purpose.

Figure 4.8 shows an Entity Relation-diagram of the database constructed for this work. The design involves the use of 4 entities: gene, transcript, protein, and domain and binary relationship between them. Each gene is characterized by its gene_ID, gene_name and description. Transcripts which belong to genes are modeled using their IDs and names. Furthermore, the domain of each transcript is stored by its Pfam_id and Smart_id. Finally, the proteins associated with each gene are stored using their IDs and names. Primary key for each entity is indicated by underline. Gene entity has a one-to-many relation to transcript and protein entities; indicated by an arrow in the figure. Each transcript must belong to a particular gene, therefore indicated by double lines in the figure. The protein to gene relation follows the same rule. Transcript and domain entities have many-to-many relation; hence, there is no arrow in their connection.

A separate but similar database is created for each of the three genomes; human, mouse and rat which are considered in this study. Each database consists 5 tables; Gene, Transcript, Domain, Pfam and Uniprot, as shown in Figure 4.9.

HumanGene table has three columns. GeneID is the primary key and TransCount, gives the number of transcripts each gene has, and the GeneName is the actual name of the gene. HumanGene table has a one to many relations with HumanTranscript

table but many-to-many relation with HumanUniprot table, since there can be more than one protein produced by a gene and a UniprotID may correspond to more than one GeneID in a phenomenon known as a Haplotypic region.

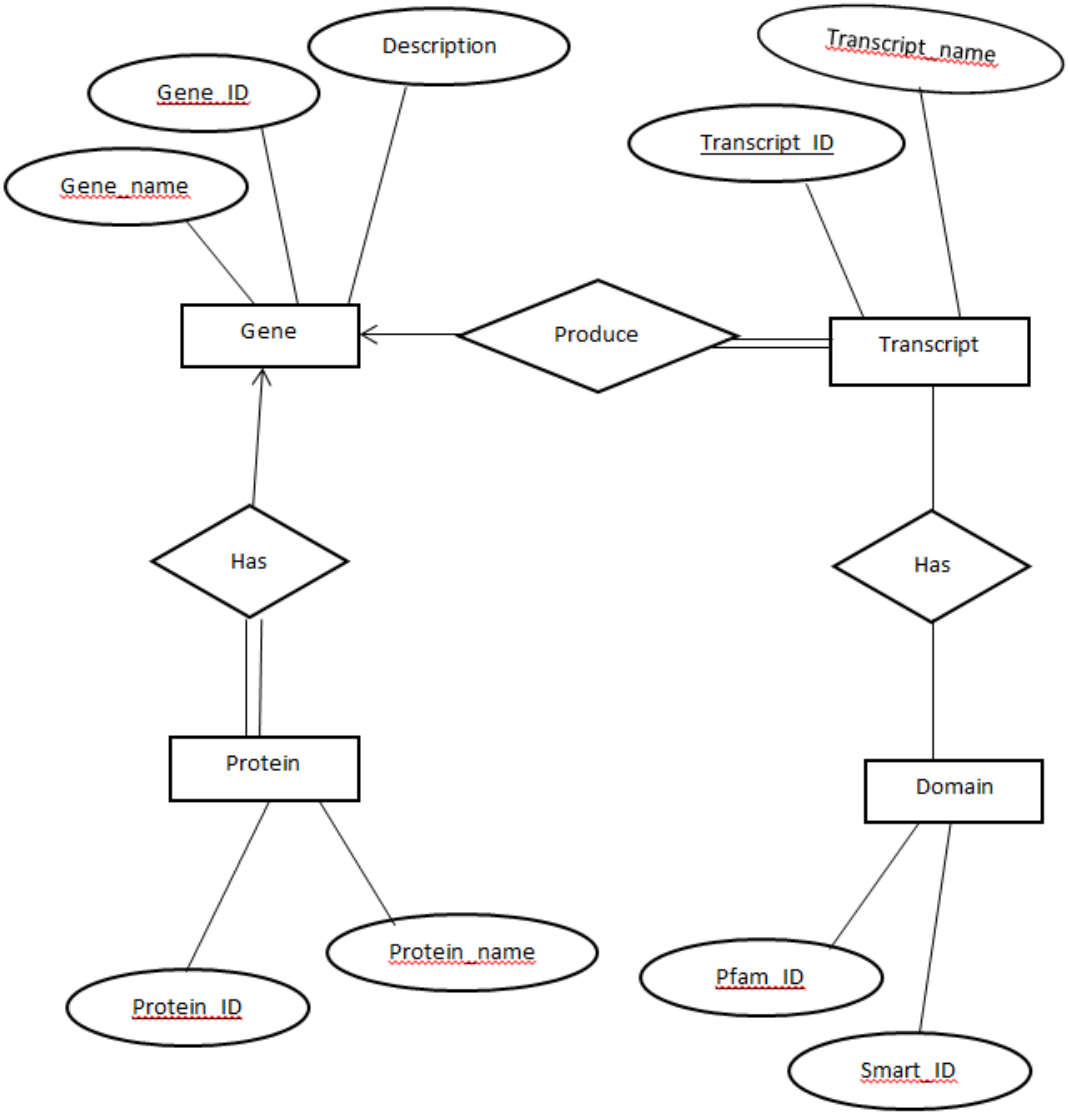


Figure 4.7: Entity Relation (E-R) diagram for the designed database showing the relationship between genes, transcripts, proteins and domains

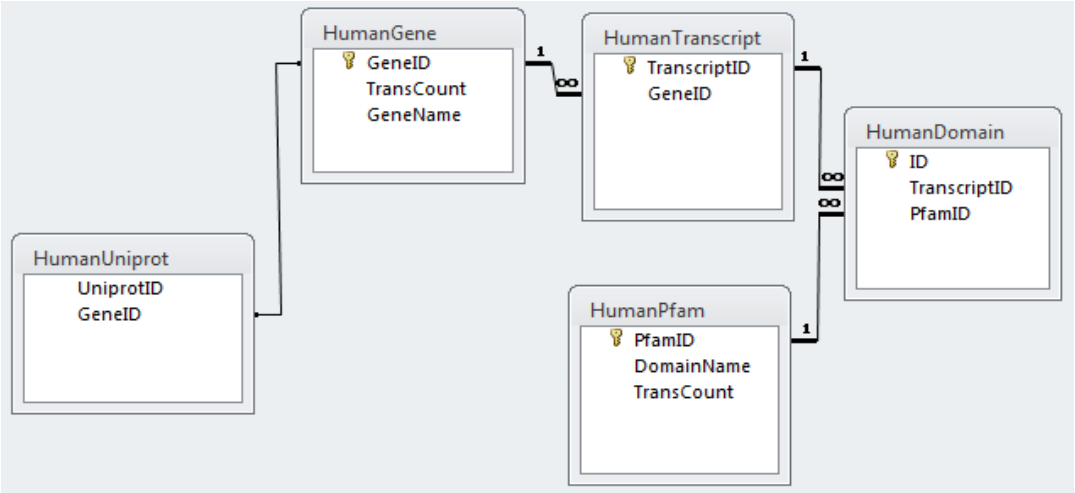


Figure 4.8: Schema diagram for Human GPCR database showing the 5 tables which constitute the database

Human Transcript table has two columns. TranscriptID is the primary key and it indicates the transcripts which correspond to the gene whose GeneID is stored in the second column. This table has a one-to-many relations with the Human Domain since it is known that there will be many transcripts associated with each domain.

HumanDomain table has three columns: ID which is the primary key for each domain, TranscriptID, indicating each transcript in a domain and PfamID which indicates the protein domain ID. The HumanDomain table has a many-to-one relation with HumanPfam table. HumanPfam table also has three columns (PfamID, DomainName and TransCount). PfamID is the primary key and indicates a protein domain in the original Pfam database. DomainName is the name of the protein domain as given in the Pfam database and TransCount is the number of Transcripts which exists in the same domain.

Similar representation of tables for both Mouse and Rat GPCRs are designed and are shown in Figures 4.10 and 4.11 respectively.

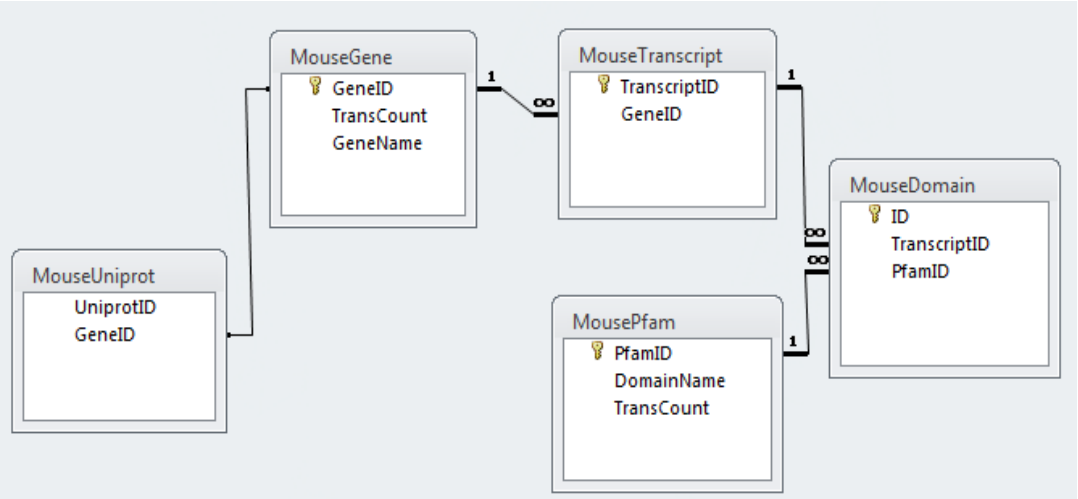


Figure 4.9: Schema diagram for Mouse GPCR database showing the 5 tables which constitute the database

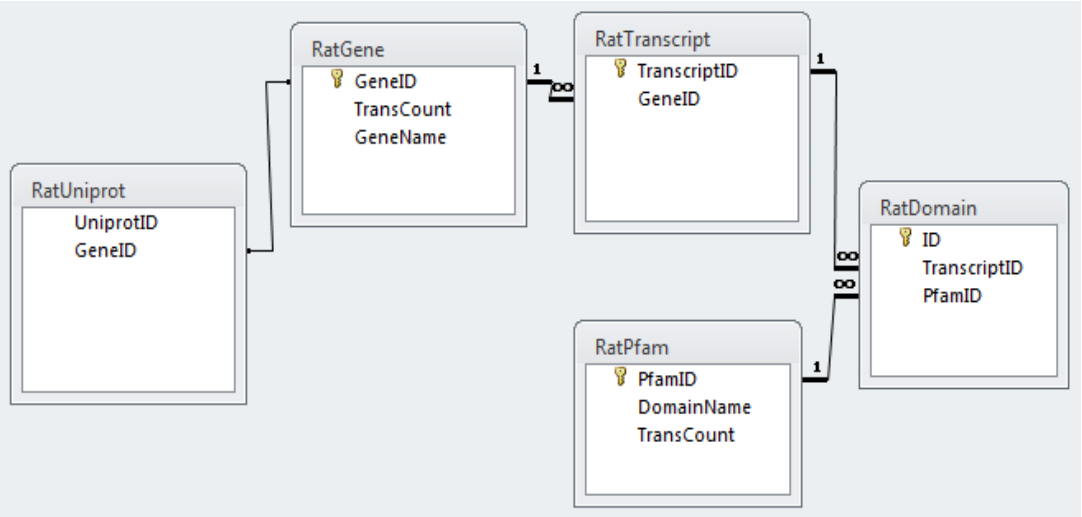


Figure 4.10: Schema diagram for Rat GPCR database showing the 5 tables which constitute the database

This database is created in order to save data retrieved from different sources (Uniprot, Biomart, and NCBI) in a single database for easy access, either for analysis, search or update. In particular, we used the database in 2 ways:

i. **Save and retrieval**; this is used for easy access to data by querying in order to retrieve data from the database depending on the intended output using

basic queries. For example, a query is written to display all mouse genes and their corresponding transcripts;

```
SELECT h.GeneID
FROM HumanGene
```

Another example query would be to count the occurrence of values (e.g. for each gene, count the number of transcripts that correspond to the gene;

```
SELECT count(distinct t.TranscriptID) as Cont
FROM h HumanGene, t HumanTranscript
WHERE t.GeneID = h.GeneID
```

ii. **Clean and search;** since the database contains many data points, there are cases where the data has to be cleaned before further analysis is done. For example, we need to find all genes where some transcript has a Pfam ID but one or more transcripts of the same gene do not have Pfam ID. The query shown below can be used for this purpose.

```
SELECT h.GeneID, count(distinct t.TranscriptID) as Cont,
h.TransCount,t.TranscriptID, count(distinct p.PfamID) as ContP
FROM HumanDomain as d, HumanPfam as p, HumanTranscript as t,
HumanGene as h
WHERE d.TranscriptID = t.TranscriptID and d.PfamID = p.PfamID and
h.GeneID = t.GeneID GROUP BY h.GeneID
HAVING Cont < h.TransCount and Cont > 1 and ContP > 1
```

The query is useful in the search for protein domain diversity (case 1) explained in section 4.4.

4.4 Hypothesis Analysis

Retrieving and saving of the necessary data for our hypothesis in this thesis is explained in Sections 4.3.1 and 4.3.2 above. The next step involves the actual analysis of the data to be explained in this section.

In order to analyze the data for protein domain diversity, firstly we define the meaning of the absence and presence of diversity. All transcripts which belong to genes in our database correspond to one or more Pfam IDs, which represent proteins these transcripts code for. It was observed that not all transcripts are included in the Pfam database or in the Smart database as well as and other related biological databases. Pfam database, however, is more comprehensive in that it includes more protein domains than the Smart database or any other databases in this domain. Therefore, the Pfam database is used in this study.

Figure 4.12 shows how absence of diversity is defined. It is done by checking if all transcripts of a particular gene code for the same number of proteins. In the figure, all three transcripts of GeneX, code for the same proteins and all these proteins, represented as PfamID1, PfamID2 and PfamID3 are in Pfam database.

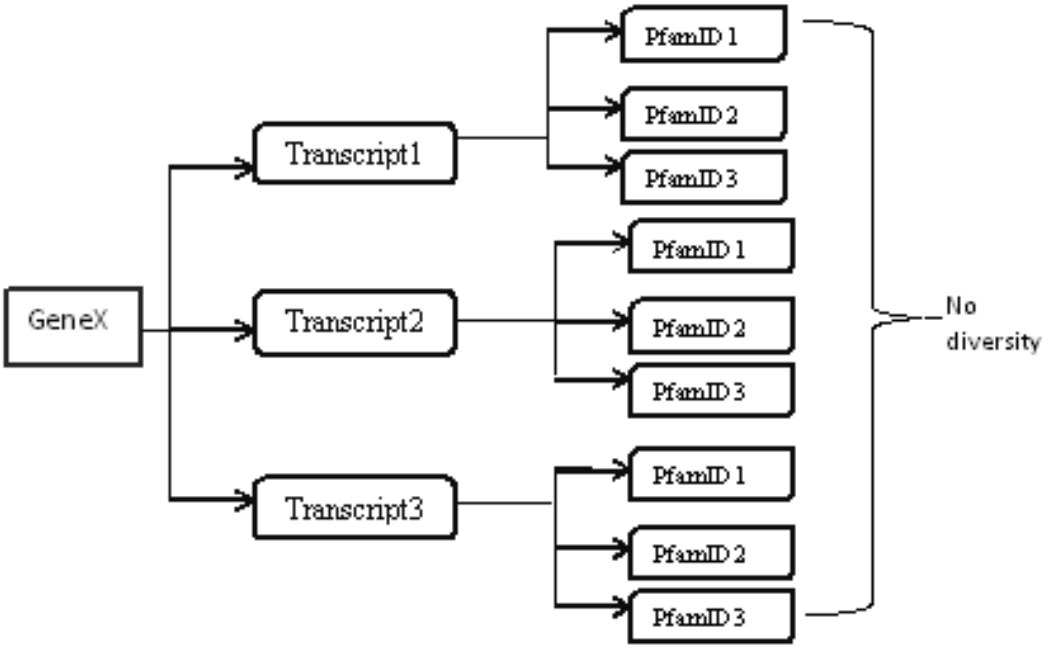


Figure 4.11: Representation of absence of protein domain diversity

The presence of protein domain diversity is defined for two different cases. In the first case, the transcripts of each gene are checked to see that all proteins coded by

the transcripts are included in Pfam; if not, this is defined as protein domain diversity as shown in Figure 4.13. In this figure, GeneX1 has three transcripts represented by Transcript1, Transcript2 and Transcript3. Transcript1 and Transcript3 do not code any protein included in Pfam while Transcript2 codes for three proteins represented by PfamID1, PfamID2 and PfamID3.

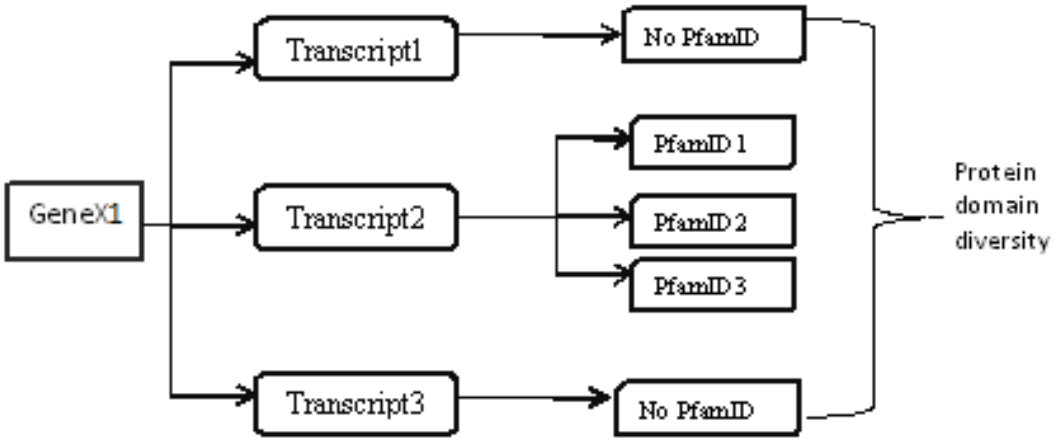


Figure 4.12: Representation of protein domain diversity (Case 1)

In the second case, all Pfam IDs are included in the Pfam database. Each gene and transcripts which belong to the particular gene is checked to see if they have a different domain from others or not. Figure 4.14 shows how the comparison is carried out. In this figure, GeneX2 has three transcripts (Transcript1, Transcript2, and Transcript3). Transcript1 codes for three proteins with Pfam IDs PfamID1, PfamID2 and PfamID3. But Transcript3 codes for only two of these proteins (PfamID2 and PfamID3). This case is defined as protein domain diversity. In fact, in this particular case, Transcript2 codes for proteins with Pfam IDs, PfamID2 and PfamID4 and Transcripts3 codes for proteins with Pfam IDs, PfamID2 and PfamID3.

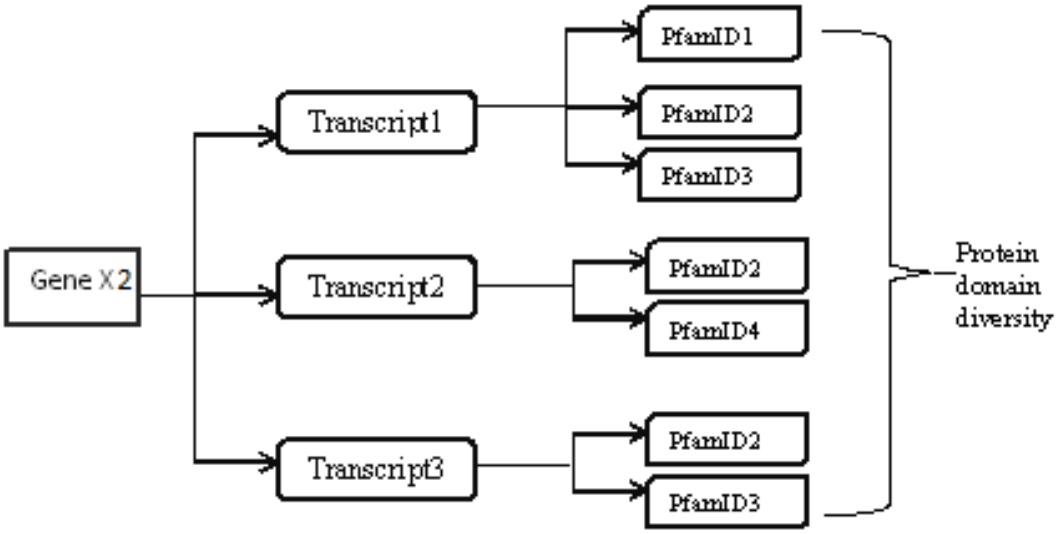


Figure 4.13: Representation of protein domain diversity (Case 2)

Chapter 5

RESULTS AND DISCUSSION

5.1 Initial Data Retrieval from UniProt

The data used is obtained from the UniProt database, which has been found to be the most comprehensive database containing GPCR proteins as mentioned in Section 4.3.1. It provides link(s) to the Ensembl database where corresponding genes are found. Table 5.1 shows the result of the query for GPCR proteins in UniProt for the three different species analyzed in this study.

Table 5.1: Number of GPCR proteins found in UniProt for different species (data retrieved in October, 2016)

Species	Number of GPCRs (Proteins in UniProt)
Human	845
Mouse	513
Rat	338

Query results in UniProt for all three species are downloaded separately in XML format. This format makes it easy for needed UniProt IDs to be extracted from the downloaded files, using a code written in Perl language. The Perl code used for this is given in APPENDIX A. APPENDICES B, C and D have the list of UniProt IDs used for obtaining corresponding GeneIDs from Biomart database to be used in the next step for human, mouse and rat, respectively.

The UniProt IDs obtained are used in Ensembl Biomart to obtain the data needed for all GPCRs in human, mouse and rat. The Biomart query results are given in Table 5.2. It was found that human genome has the highest number of GPCRs among the three genomes analyzed with 920 genes found in Biomart database, followed by mouse.

Table 5.2: Number of GPCR genes in different species as found in Biomart (data retrieved in October, 2016)

Species	Number of GPCRs
Human	920
Mouse	473
Rat	205

5.2 GPCRs with multiple transcripts

The result from Table 5.2 is further analyzed to retrieve data for only genes that have multiple transcripts and are protein coding. It was found that majority of GPCRs have only one transcript with about 64% in human, about 53% in mouse and about 83% in rat. These results are shown in Table 5.3 and Figure 5.1.

Further analysis is shown in Table 5.4, where different number of genes with different amount of transcripts is tabulated for the species being analyzed. A query is written to check our database for genes with a particular number of transcripts. This query searches for genes with “num” transcripts, where num is a variable for the number of transcripts GPCRs to be shown should code for.

```
CREATE PROCEDURE `transcriptNumber`(IN `num` INT UNSIGNED)
NOT DETERMINISTIC READS SQL DATA SQL SECURITY INVOKER
SELECT * FROM HumanGene WHERE TransCount = num
```

Table 5.3: Number of GPCRs with single transcript and those with multiple transcripts

Species	Number of GPCRs with one transcript	Percentage of GPCRs with one transcript	Number of GPCRs with multiple transcripts	Percentage of GPCRs with multiple transcripts
Human	585	63.59%	335	36.41%
Mouse	249	52.64%	224	47.36%
Rat	171	83.41%	34	16.59%

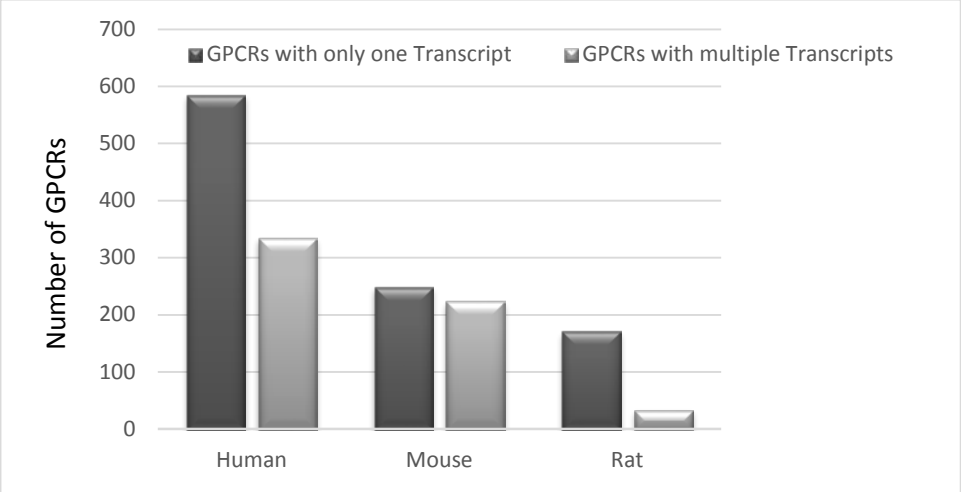


Figure 5.1: GPCRs with single transcripts versus those with multiple transcripts

This query was executed for all possible number of transcripts and the results are shown in Table 5.4 and Figure 5.2. The results show that for the three species, a total of two transcripts are the most common for genes with multiple transcripts. The highest transcript number per gene in human is seventy seven (77), thirty five (35) in mouse and twenty (20) in rat.

Table 5.4: Number of Transcripts per GPCR gene

Transcripts count per gene	Occurrence in human GPCR genes	Occurrence in mouse GPCR genes	Occurrence in rat GPCR genes
1	585	249	171
2	109	87	23
3	64	43	6
4	32	28	2
5	31	12	0
6	14	21	1
7	26	7	0
8	9	8	0
9	6	6	0
10	5	1	0
11	8	4	0
12	4	1	1
13	9	1	0
14	5	0	0
15	4	1	0
16	2	1	0
17	1	0	0
18	1	0	0
20	1	1	1
21	0	0	0
22	1	0	0
23	1	0	0
31	0	1	0
35	0	1	0
77	1	0	0

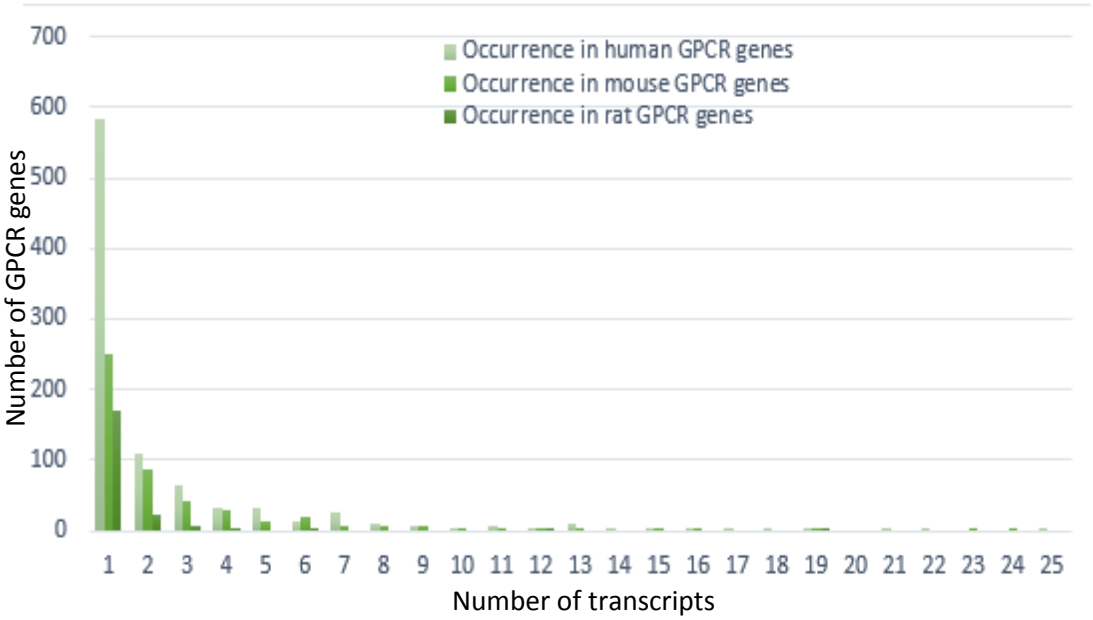


Figure 5.2: Transcripts per GPCR gene

5.3 GPCRs with protein-coding transcripts

In the next step, GPCRs that are not protein coding are eliminated from the data. Majority of GPCRs in our data were found to be protein-coding in all three species; 97.28% in human, 100% in both mouse and rat as shown in Table 5.5.

Table 5.5: Protein coding and non-protein coding GPCR genes

Species	Number of protein-coding GPCRs	Percentage of protein-coding GPCRs	Number of non-protein-coding GPCRs	Percentage of non-protein-coding GPCRs
Human	895	97.28%	25	2.72%
Mouse	473	100%	0	0.00%
Rat	205	100%	0	0.00%

Considering GPCRs that have multiple transcripts and are protein-coding, it was found that majority or all GPCRs with multiple transcripts are protein coding; almost 100% in human, and a 100% in both mouse and rat as shown in Table 5.6. This part of the result (that is GPCRs that are protein-coding and with multiple transcripts) formed the main data in the database that was further analyzed.

Table 5.6: Protein coding and non-protein coding GPCR genes with multiple Transcripts

Species	Number of protein-coding GPCRs with multiple Transcripts	Percentage of protein-coding GPCRs with multiple Transcripts	Number of non-protein coding GPCRs with multiple Transcripts	Percentage of non-protein coding GPCRs with multiple Transcripts
Human	334	99.70%	1	0.30%
Mouse	224	100%	0	0.00%
Rat	34	100%	0	0.00%

Protein domains are of major importance in this work. Therefore, we checked how many domains can be found in the protein coded by the analyzed transcripts. First, we distinguished between transcripts coding for single domains and those coding for multiple domains in all three species. The results as given in Table 5.7 show that there are more transcripts coding for one domain than those coding for multiple domains across all three species; 72% in human, 79% in mouse and 47% in rat.

Table 5.7: Number of transcripts coding for a single domain and those coding for multiple domains

Species	Number of Transcripts coding for only one Domain	Percentage of Transcripts coding for only one Domain	Number of Transcripts coding for multiple Domain	Percentage of Transcripts coding for multiple Domain
Human	546	72.3%	209	27.70%
Mouse	328	78.7%	89	21.30%
Rat	27	46.6%	31	53.40%

In addition to checking for transcripts coding for single domains and those with multiple domains, transcripts having certain number of domains were check and the results are given in Table 5.8. The following sql query is used for human;

```
CREATE PROCEDURE `domainNumber` (IN `num1` INT UNSIGNED) NOT
DETERMINISTIC NO SQL SQL SECURITY DEFINER
SELECT g.GeneID, t.TranscriptID, COUNT (DISTINCT d.PfamID) as
DomainCount
FROM HumanGene as g, Humantranscript as t, Humandomain as d
WHERE g.GeneID = t.GeneID and t.TranscriptID = d.TranscriptID
GROUP BY g.GeneID, t.TranscriptID
HAVING DomainCount = num1
```

where num1 is the variable for number of domains (from 1 to 8) as seen in the first column of Table 5.8.

Similar queries were carried out for mouse and rat and results are represented in Table 5.8. Most GPCR gene transcripts code for fewer amount of domains across all three species. This is shown in Figure 5.3, where one domain is coded by most transcripts, followed by two domains; decreasing in that order. Queries for mouse and rat species are shown in Appendix E and F respectively.

Table 5.8: Total number of domains coded by each GPCR transcript

Number of domains	Human GPCR transcripts	Mouse GPCR transcripts	Rat GPCR transcripts
1	546	328	27
2	106	48	20
3	60	19	7
4	23	9	2
5	4	5	1
6	0	1	0
7	15	7	1
8	1	0	0

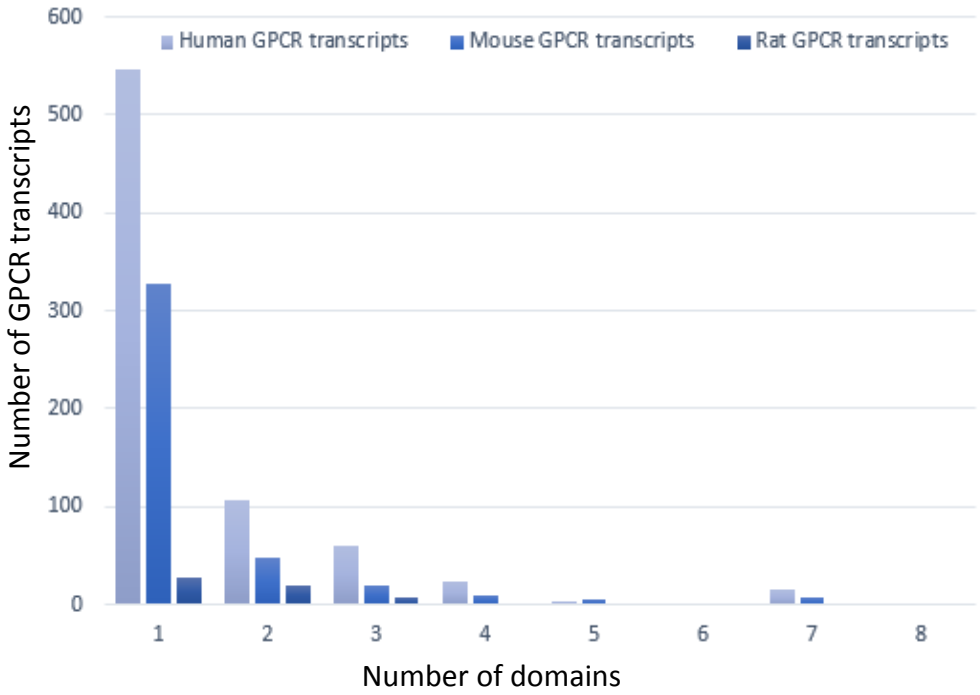


Figure 5.3: Protein domains versus GPCR transcripts

Table 5.9 shows the Biomart result after applying the filters necessary for this work as described in Chapter 4. A in the figure shows “Ensembl Gene IDs” used in Ensembl to identify each unique gene. B shows transcripts of genes represented in A, C gives short description of each gene, while D indicates the number of transcripts

that exists in each gene. E shows the biotype of each transcript; which is essential to this study because this work is only interested in “protein coding” biotype. Columns showing “Pfam ID” and “UniProtID” which are links to these genes in external databases, Pfam and UniProt respectively, as well as gene name, are added later so as to have comprehensive database.

Further analysis is based on the results obtained at this point.

Table 5.9: Biomart result for human GPCRs (data retrieved in December, 2016)

A	B	C	D	E
Ensembl Gene ID	Ensembl Transcript ID	Description	Transcript count	Transcript type
ENSG00000270898	ENST00000406625	GPR75-ASB3 readthrough [Source:HGNC Symbol;Acc:HGNC:40043]	3	protein_coding
ENSG00000158301	ENST00000332262	G protein-coupled receptor associated sorting protein 2 [Source:HGNC Symbol;Acc:HGNC:25169]	5	protein_coding
ENSG00000158301	ENST00000543253	G protein-coupled receptor associated sorting protein 2 [Source:HGNC Symbol;Acc:HGNC:25169]	5	protein_coding
ENSG00000158301	ENST00000535209	G protein-coupled receptor associated sorting protein 2 [Source:HGNC Symbol;Acc:HGNC:25169]	5	protein_coding
ENSG00000173698	ENST00000379873	adhesion G protein-coupled receptor G2 [Source:HGNC Symbol;Acc:HGNC:4516]	11	protein_coding
ENSG00000173698	ENST00000357991	adhesion G protein-coupled receptor G2 [Source:HGNC Symbol;Acc:HGNC:4516]	11	protein_coding
ENSG00000173698	ENST00000357544	adhesion G protein-coupled receptor G2 [Source:HGNC Symbol;Acc:HGNC:4516]	11	protein_coding
ENSG00000173698	ENST00000379869	adhesion G protein-coupled receptor G2 [Source:HGNC Symbol;Acc:HGNC:4516]	11	protein_coding
ENSG00000173698	ENST00000360279	adhesion G protein-coupled receptor G2 [Source:HGNC Symbol;Acc:HGNC:4516]	11	protein_coding
ENSG00000173698	ENST00000356606	adhesion G protein-coupled receptor G2 [Source:HGNC Symbol;Acc:HGNC:4516]	11	protein_coding
ENSG00000173698	ENST00000340581	adhesion G protein-coupled receptor G2 [Source:HGNC Symbol;Acc:HGNC:4516]	11	protein_coding
ENSG00000173698	ENST00000379876	adhesion G protein-coupled receptor G2 [Source:HGNC Symbol;Acc:HGNC:4516]	11	protein_coding
ENSG00000173698	ENST00000379878	adhesion G protein-coupled receptor G2 [Source:HGNC Symbol;Acc:HGNC:4516]	11	protein_coding
ENSG00000173698	ENST00000354791	adhesion G protein-coupled receptor G2 [Source:HGNC Symbol;Acc:HGNC:4516]	11	protein_coding
ENSG00000105808	ENST00000262940	RAS p21 protein activator 4 [Source:HGNC Symbol;Acc:HGNC:23181]	17	protein_coding
ENSG00000105808	ENST00000461209	RAS p21 protein activator 4 [Source:HGNC Symbol;Acc:HGNC:23181]	17	protein_coding
ENSG00000105808	ENST00000449970	RAS p21 protein activator 4 [Source:HGNC Symbol;Acc:HGNC:23181]	17	protein_coding
ENSG00000105808	ENST00000462172	RAS p21 protein activator 4 [Source:HGNC Symbol;Acc:HGNC:23181]	17	protein_coding
ENSG00000105808	ENST00000521397	RAS p21 protein activator 4 [Source:HGNC Symbol;Acc:HGNC:23181]	17	protein_coding
ENSG00000105808	ENST00000522801	RAS p21 protein activator 4 [Source:HGNC Symbol;Acc:HGNC:23181]	17	protein_coding
ENSG00000105808	ENST00000520042	RAS p21 protein activator 4 [Source:HGNC Symbol;Acc:HGNC:23181]	17	protein_coding
ENSG00000105808	ENST00000521076	RAS p21 protein activator 4 [Source:HGNC Symbol;Acc:HGNC:23181]	17	protein_coding
ENSG00000105808	ENST00000538869	RAS p21 protein activator 4 [Source:HGNC Symbol;Acc:HGNC:23181]	17	protein_coding
ENSG00000139679	ENST00000378434	lysophosphatidic acid receptor 6 [Source:HGNC Symbol;Acc:HGNC:15520]	9	protein_coding

5.4 Analysis of final data

5.4.1 Transcript diversity in human GPCRs

Diversity in domains coded by one transcript compared to another of the same gene is checked through the methodology explained in Chapter 4. There are different cases of domain diversity. In case 1, all transcripts per gene are screened for domains in Pfam. Transcripts which do not code for any known Pfam domains are noted; as well as transcripts which code for known domains with associated Pfam IDs are identified

as described in Section 4.2. In this case, diversity is identified based on the existence or the absence of domains coded by different transcripts of the same gene.

For human, the following sql query is used for this first case;

```
SELECT h.GeneID, count (distinct t.TranscriptID) as Cont, h.TransCount
FROM Humandomain as d, Humanpfam as s, Humantranscript as t, Humangene as h
WHERE d.TranscriptID = t.TranscriptID and d.PfamID = s.PfamID and h.GeneID = t.GeneID
GROUP BY h.GeneID
HAVING Cont < h.TransCount
```

Case 2 describes genes with multiple transcripts, all of which code for domains listed in Pfam. In this case, one-to-one comparisons of all domains in all transcripts in a gene are performed to determine the differences between them. The query used for this process is as follows;

```
SELECT h.GeneID, t.TranscriptID, count(distinct t.TranscriptID) as Cont, h.TransCount, count(distinct p.PfamID) as ContP
FROM Humandomain as d, Humanpfam as p, Humantranscript as t, Humangene as h
WHERE d.TranscriptID = t.TranscriptID and d.PfamID = p.PfamID and h.GeneID = t.GeneID
GROUP BY  h.GeneID
HAVING Cont < h.TransCount and Cont > 1  and ContP > 1
AND Cont * ContP != h.TransCount
```

The above queries result in 275 GPCRs as genes with domain diversity, which represents 82.83% of the total number of GPCR genes in human. This analysis is based on those genes with multiple transcripts, protein coding and covered by Pfam. The results are presented in Table 5.10 and the list of the GPCRs with transcript domain diversity is given in Appendix G.

Table 5.10: Human GPCR protein domain diversity

Protein-coding genes with more than one transcripts	334
Genes with Pfam IDs	332
Genes with Transcript/Domain Diversity using Pfam IDs	275
Percentage of Transcript/Domain Diversity using Pfam IDs	82.83%

An example of a GPCR gene with domain diversity in human is **CRHR1**. CRHR1 corresponds to the Ensembl ID: ENSG00000120088. It is described as corticotropin releasing hormone receptor 1 and it has 14 transcripts, 8 of which are protein coding. Table 5.11 is a screenshot adapted from Ensembl showing a list of transcripts of this gene and their corresponding protein size and biotype.

Table 5.11: List of transcripts of CRHR1 gene (data retrieved in November, 2016)

A	B	C	D	E	F	G	H
Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq
CRHR1-002	ENST00000314537.9	2535	415aa	Protein coding	CCDS42350	P34998	NM_004382 NP_004373
CRHR1-202	ENST00000339069.9	2490	314aa	Protein coding	CCDS77049	B3TIK8	NM_001303016 NM_001303020 NP_001289945 NP_001289949
CRHR1-001	ENST00000398285.7	2399	444aa	Protein coding	CCDS45712	P34998	NM_001145146 NP_001138618
CRHR1-004	ENST00000577353.5	1206	401aa	Protein coding	CCDS45713	P34998	NM_001145148 NP_001138620
CRHR1-003	ENST00000352855.9	1146	375aa	Protein coding	CCDS45714	P34998	NM_001145147 NP_001138619
CRHR1-201	ENST00000293493.11	2621	430aa	Protein coding	-	A0A0A0MQZ1	-
CRHR1-203	ENST00000619154.4	2370	154aa	Protein coding	-	K9J956	-
CRHR1-009	ENST00000580876.5	223	75aa	Protein coding	-	J3QKP8	-
CRHR1-005	ENST00000347197.9	2462	145aa	Nonsense mediated decay	-	J9JIC6	-
CRHR1-008	ENST00000535778.2	1272	52aa	Nonsense mediated decay	-	H0YFR5	-
CRHR1-010	ENST00000583888.1	777	152aa	Nonsense mediated decay	-	J3KSM0	-
CRHR1-006	ENST00000580955.5	479	93aa	Nonsense mediated decay	-	J3QQR1	-
CRHR1-007	ENST00000582766.5	828	No protein	Retained intron	-	-	-
CRHR1-011	ENST00000581479.1	606	No protein	Retained intron	-	-	-

A in Table 5.11 shows “Name” of each transcript of the gene, which is a combination of the transcript’s gene name and its serial number. B shows the identity number of each transcript as shown in Ensembl Biomart in Section 5.5, though with a sub digit which indicates transcript version number. D shows the lengths of proteins coded by the transcript in amino acid. This column also serves as link to “Protein summary”, where the domains are shown graphically as can be seen in Figures 5.5, 5.6 and 5.7. E shows the biotype; that is the type of the transcripts while F, G and H are links to other databases, such as UniProt, where more details about the transcripts can be found.

Figure 5.4 is a graphical representation of Table 5.11, where each transcript in the gene is represented by a line; red line for protein coding transcripts and blue line for non-protein coding transcripts. More information about these transcripts (such as protein coded by the transcripts, transcripts type, amino acids, and gene alleles) can be obtained by clicking on the lines in the figure.

Figures 5.5, 5.6 and 5.7 show three transcript domains of CRHR1. ENST00000314537 (Figure 5.5) and ENST00000398285 (Figure 5.6) both have two domains, which correspond to PF02793 and PF00002 in Pfam, while ENST00000339069 (Figure 4.7) codes for only one domain; PF00002. The figures also show different representations of protein domain and links of other biological databases similar to Pfam, such as Print domain, Smart domains and Superfamily domains, which are usually not quiet consistent with Pfam that is used in this study. They also show sequence variants of the proteins among others.

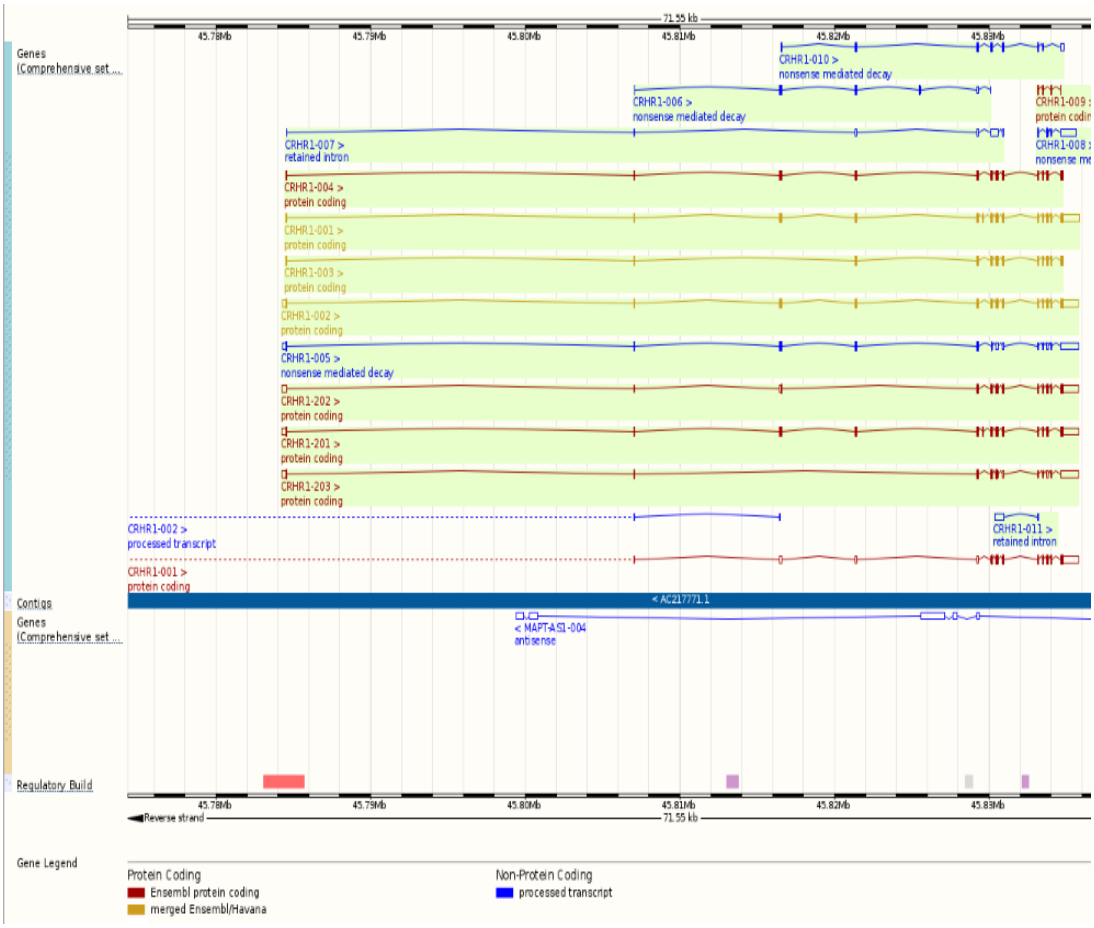


Figure 5.4: Graphical representation of transcripts in CRHR1 as shown in Ensembl (data retrieved in November, 2016)

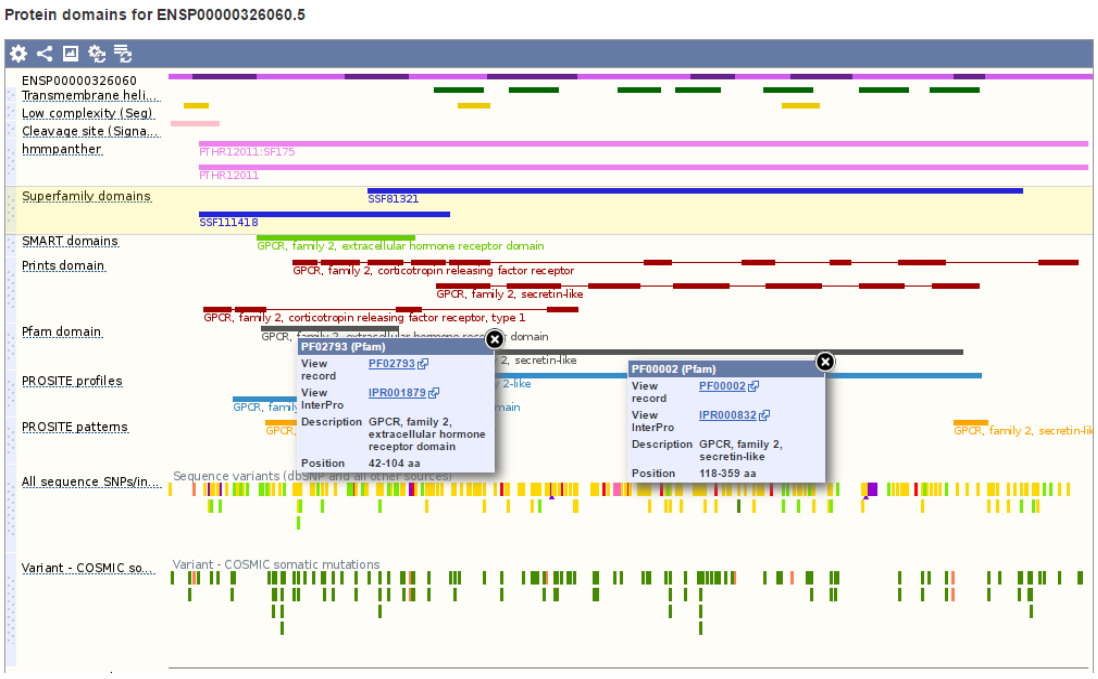


Figure 5.5: Summary for transcript: ENST00000314537.9, Gene: CRHR1 (data retrieved in November, 2016)

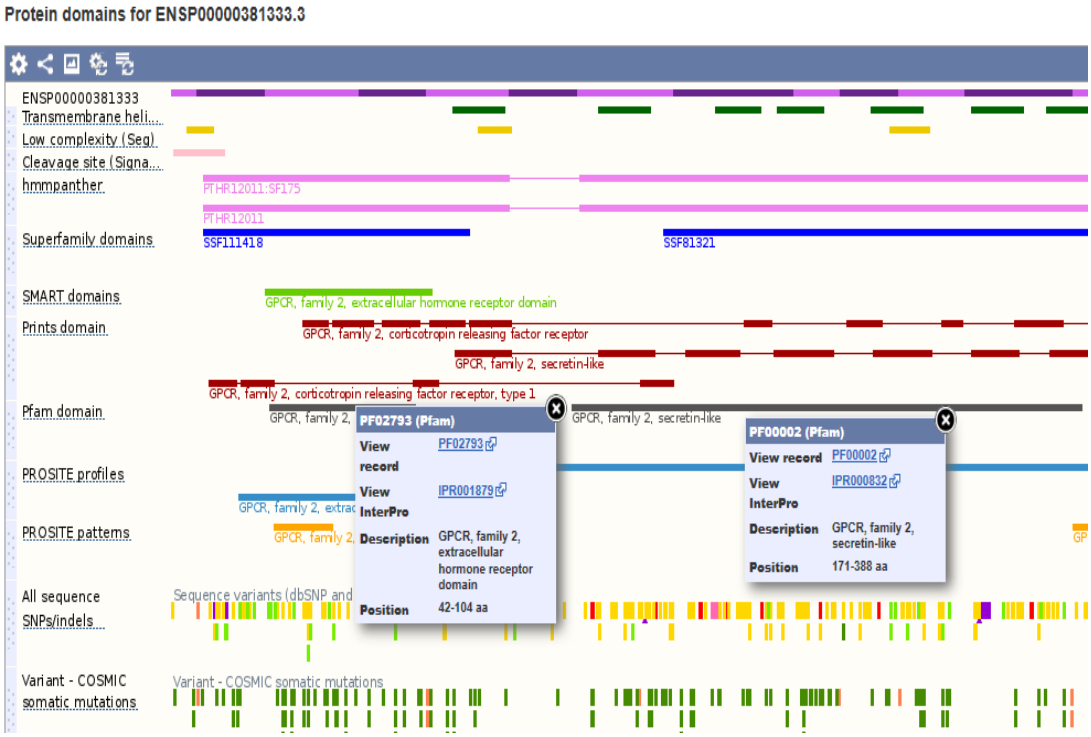


Figure 5.6: Summary for transcript: ENST00000398285.7, Gene: CRHR1 (data retrieved in November, 2016)

Protein summary

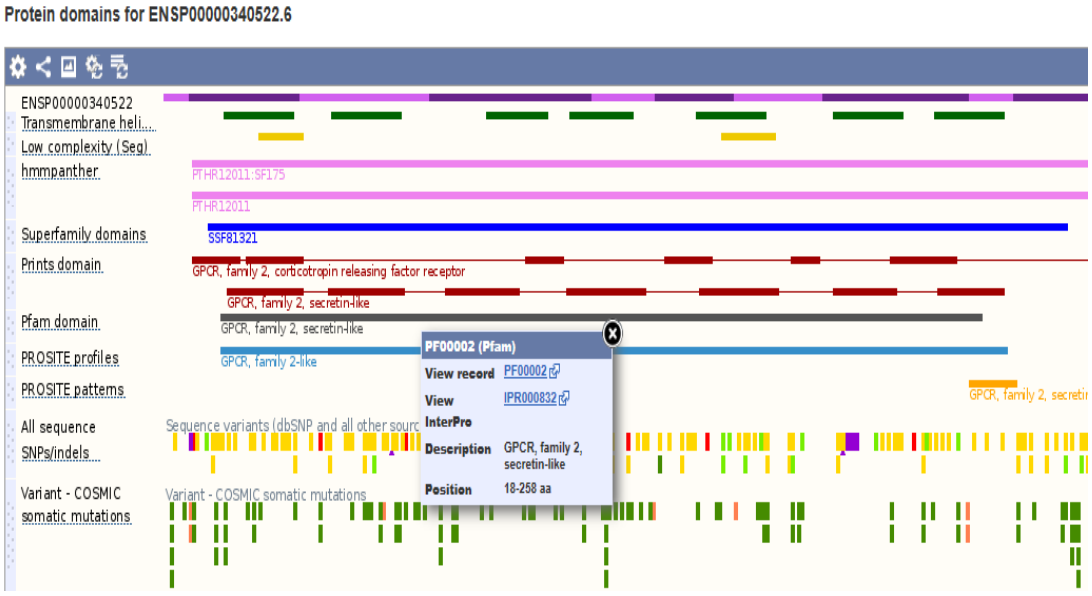


Figure 5.7: Summary for transcript: ENST00000339069.9, Gene: CRHR1 (data retrieved in November, 2016)

5.4.2 Transcript diversity in mouse GPCRs

For mouse, as in the case of human, the following sql query is used for case 1, as described in Section 5.3;

```
SELECT h.GeneID, count (distinct t.TranscriptID) as Cont, h.TransCount
FROM Mousedomain as d, Mousepfam as s, Mousetranscript as t, Mousegene as h
WHERE d.TranscriptID = t.TranscriptID and d.PfamID = s.PfamID and h.GeneID = t.GeneID
GROUP BY h.GeneID
HAVING Cont < h.TransCount
```

The following query is used for Case 2 as described in Section 5.3;

```
SELECT h.GeneID, t.TranscriptID, count (distinct t.TranscriptID) as Cont, h.TransCount, count (distinct p.PfamID) as ContP
FROM Mousedomain as d, Mousepfam as p, Mousetranscript as t, Mousegene as h
WHERE d.TranscriptID = t.TranscriptID and d.PfamID = p.PfamID and h.GeneID = t.GeneID
GROUP BY h.GeneID
HAVING Cont < h.TransCount and Cont > 1 and ContP > 1
AND Cont * ContP != h.TransCount
```

The result of using the above queries gives 180 GPCRs as genes with domain diversity in mouse, which represents 81% of the total number of GPCR genes. This analysis is based on those with multiple transcripts, protein coding and covered by Pfam. The results are presented in Table 5.12, and the list of these GPCRs is given in Appendix H.

An example of a GPCR gene with transcripts diversity in mouse is **Adgrl1**. Adgrl1 corresponds to the Ensembl ID: ENSMUSG00000013033 described as an adhesion G protein-coupled receptor L1 and has proteins that correspond to the UniProtKB identifier: Q80TR1. It has 10 transcripts, 6 of which are protein coding. Table 5.13 is a screenshot adapted from Ensembl showing a list of transcripts present in this gene

and their corresponding protein and biotype as well as other columns as explained in sub-section 5.4.1.

Table 5.12: Mouse GPCR protein domain diversity

Protein-coding genes with more than one transcripts	224
Genes with Pfam ID	222
Genes with Transcript/Domain Diversity using Pfam ID	180
Percentage of Transcript/Domain Diversity using Pfam ID	81.08%

Table 5.13: List of transcripts of Adgrl1 gene (data retrieved in November, 2016)

A	B	D	E	F	G	H
Name	Transcript ID	bp	Protein	Biotype	UniProt	RefSeq
Adgrl1-001	ENSMUST00000141158.7	8181	1466aa	Protein coding	Q80TR1	NM_181039 NP_851382
Adgrl1-005	ENSMUST00000152978.7	5734	1516aa	Protein coding	E9Q3V9	-
Adgrl1-004	ENSMUST00000132500.7	5719	1511aa	Protein coding	E9Q9Q9	-
Adgrl1-003	ENSMUST00000045393.14	5643	1471aa	Protein coding	H7BX15	-
Adgrl1-006	ENSMUST00000131717.1	5152	1295aa	Protein coding	Q80TR1	-
Adgrl1-010	ENSMUST00000131018.1	639	213aa	Protein coding	F6YP92	-
Adgrl1-002	ENSMUST00000124355.7	5935	141aa	Nonsense mediated decay	D6RI60	-
Adgrl1-007	ENSMUST00000139575.1	756	No protein	Processed transcript	-	-
Adgrl1-008	ENSMUST00000150674.1	909	No protein	Retained intron	-	-
Adgrl1-009	ENSMUST00000141661.1	440	No protein	Retained intron	-	-

Figure 5.8 is a graphical representation of Table 5.13 where each transcript in the gene is represented by a line; red line for protein coding transcripts and blue line for

non-protein coding transcripts. It contains all other features similar to the example given in human, and described in Section 5.4.1.

Table 5.12 includes a column of proteins (E) corresponding to a transcript and a link to protein summary where the domains are shown graphically as can be seen in Figures 5.10, 5.11 and 5.12.

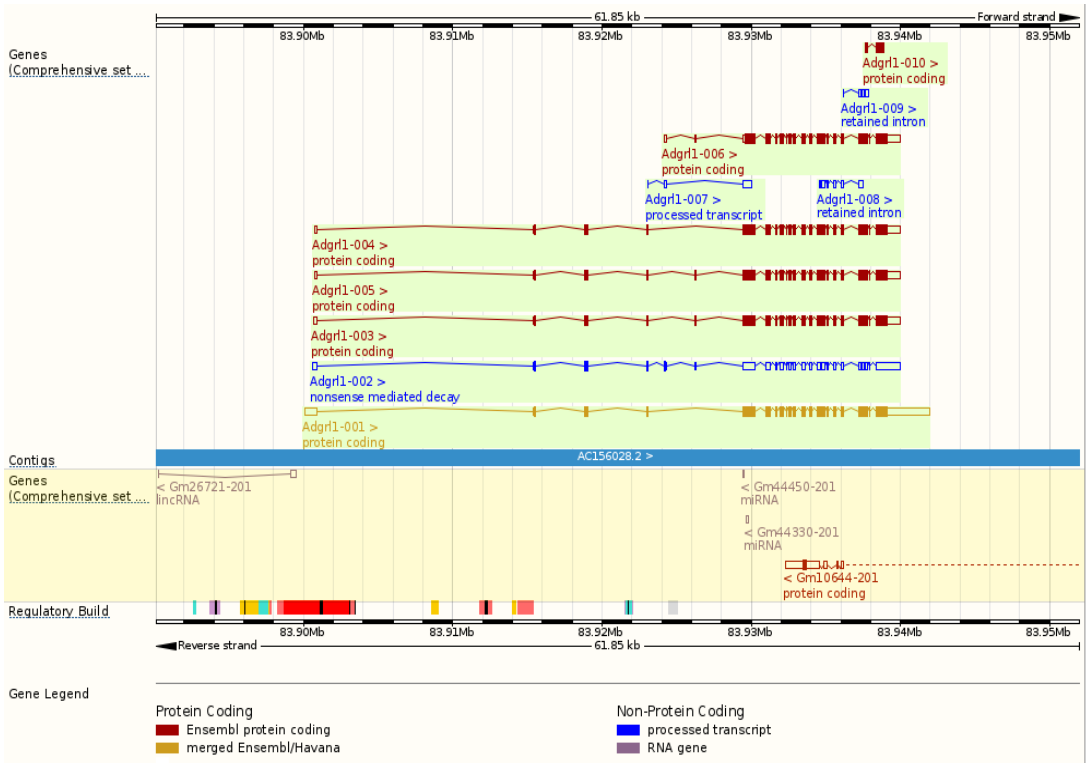


Figure 5.8: Graphical representation of transcripts in Adgrl1 as shown in Ensembl (data retrieved in November, 2016)

Figures 5.9, 5.10 and 5.11 show three transcript domains of Adgrl1. ENSMUST00000141158 in Figure 5.9 has three domains (PF02191, PF02793, PF02354), which is the same for 4 other transcripts, ENSMUST00000132500, ENSMUST00000045393, ENSMUST00000131717 and ENSMUST00000152978 but ENSMUST00000131018 (Fig. 5.10) and ENSMUST00000124355 (Fig 5.11) has only one domain each; PF02354 and PF02140 respectively. The figures also

show different representations of protein domain and links of other biological databases similar to Pfam as described in Section 5.4.1.

Figure 5.9: Summary for transcript: ENSMUST00000141158, gene: Adgrl1 (data retrieved in November, 2016)

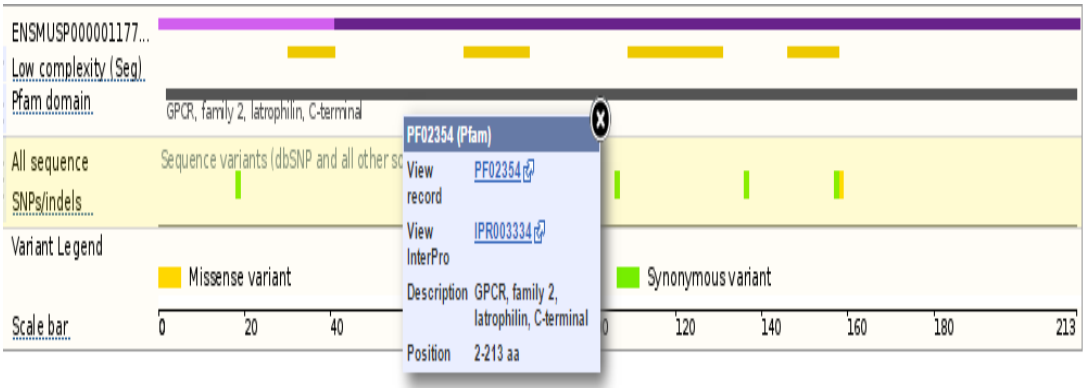


Figure 5.10: Summary for transcript: ENSMUST00000131018, gene: Adgrl1 (data retrieved in November, 2016)

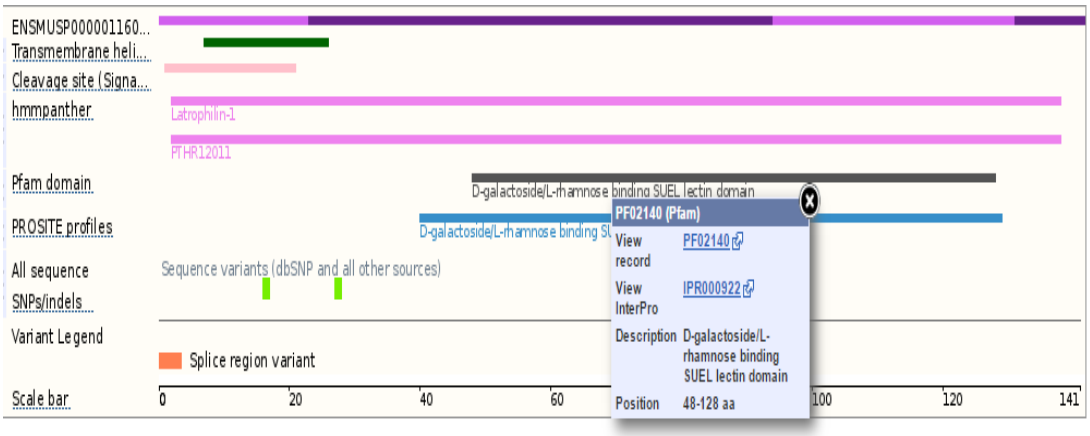


Figure 5.11: Summary for transcript: ENSMUST00000124355, gene: Adgrl1 (data retrieved in November, 2016)

5.4.3 Transcript diversity in rat GPCRs

For rat, as in the cases of human and mouse, the following sql query is used for Case 1 as described in Section 5.3;

```
SELECT h.GeneID, count(distinct t.TranscriptID) as Cont, h.TransCount
FROM Ratdomain as d, Ratpfam as s, Rattranscript as t, Ratgene as h
WHERE d.TranscriptID = t.TranscriptID and d.PfamID = s.PfamID and
h.GeneID = t.GeneID
GROUP BY  h.GeneID
HAVING Cont < h.TransCount
```

Case 2 is done by doing a one-to-one comparison of all domains in all transcripts in a gene to determine the differences between them as described in Section 5.3. The query used for this process is as follows;

```
SELECT h.GeneID, t.TranscriptID, count(distinct t.TranscriptID) as Cont,
h.TransCount, count(distinct p.PfamID) as ContP
FROM Ratdomain as d, Ratpfam as p, Rattranscript as t, Ratgene as h
WHERE d.TranscriptID = t.TranscriptID and d.PfamID = p.PfamID and
h.GeneID = t.GeneID
GROUP BY h.GeneID
HAVING Cont < h.TransCount and Cont > 1  and ContP > 1
AND Cont * ContP != h.TransCount
```

The result of using the above queries gives 22 GPCRs as genes with domain diversity in rat, which represents about sixty five percent of the total number of genes. This analysis is based on those with multiple transcripts, protein coding and covered by Pfam. The results are presented in Table 5.14; the list of the GPCRs with transcript domain diversity is given in Appendix I.

Table 5.14: Rat GPCR protein domain diversity

Protein-coding genes with more than one transcripts	34
Genes with Pfam ID	34
Genes with Transcript/Domain Diversity using Pfam ID	22
Percentage of Transcript/Domain Diversity using Pfam ID	64.71%

An example of a GPCR gene with transcripts diversity in rat is **Avpr1a**. It corresponds to the Ensembl ID: ENSRNOG00000004400, described as an arginine vasopressin receptor 1A and has proteins that correspond to the UniProtKB identifier: P30560. It has 2 transcripts, both of which are protein coding. Table 5.15 is a screenshot adapted from Ensembl showing a list of transcripts present in this gene and their corresponding protein and biotype.

Table 5.15: List of transcripts of Avpr1a gene (data retrieved in November, 2016)

A	B	D	E	F	G	H
Name	Transcript ID	bp	Protein	Biotype	RefSeq	Flags
Avpr1a-201	ENSRNOT00000005829.5	1900	424aa	Protein coding	NM_053019 NP_444178	APPRIS P1
Avpr1a-202	ENSRNOT00000087045.1	1599	394aa	Protein coding	-	

Figure 5.12 is a graphical representation of Table 5.14 where each transcript in the gene is represented by a line; red line for protein coding transcripts and blue line for

non-protein coding transcripts although there are only red lines in this figure because, there are only two transcripts of this gene, both of which are protein coding. The figure contains all other features similar to the example given in human, and described in Section 5.4.1.

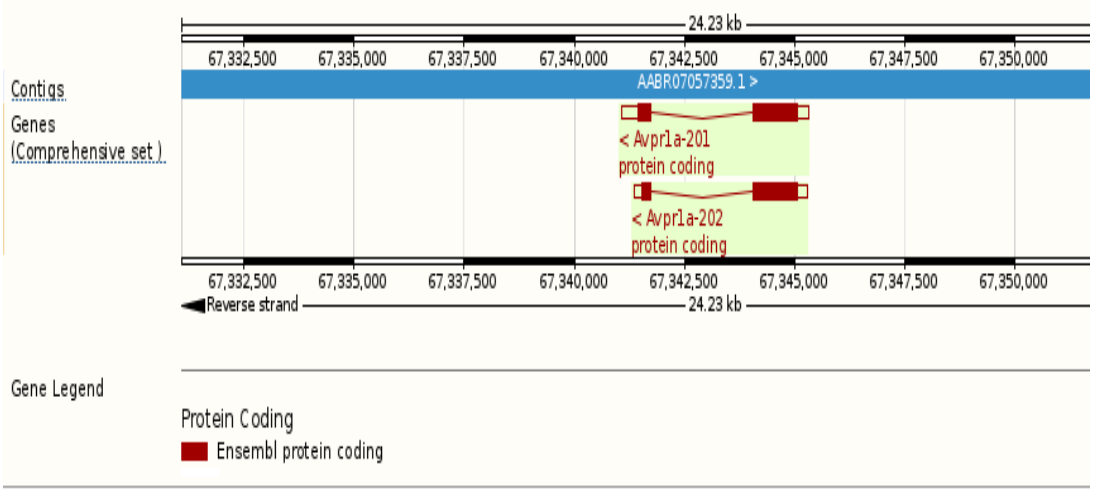


Figure 5.12: Graphical representation of transcripts in Avpr1a as shown in Ensembl (data retrieved in November, 2016).

Table 5.13 includes a column of proteins corresponding to a transcript and a link to protein summary where the domains are shown graphically as can be seen in Figures 5.9, 5.10 and 5.11.

Figures 5.14 and 5.15 show two transcript domains of Avpr1a. ENSRNOT00000005829 (Fig. 5.14) has two domains corresponding to Pfam IDs PF00001 and PF08983, while ENSRNOT00000087045 (Fig. 5.15) has only one of these domains corresponding to Pfam ID, PF00001. The figures also show different representations of protein domain and links of other biological databases similar to Pfam as described in Section 5.4.1.

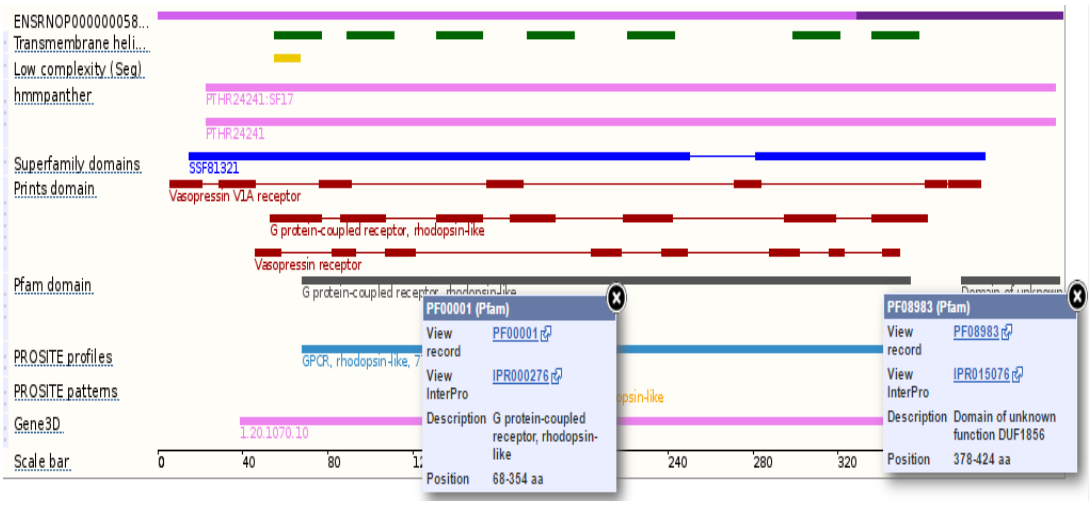


Figure 5.13: Summary for transcript: ENSRNOT00000005829, gene: Avpr1a (data retrieved in November, 2016)

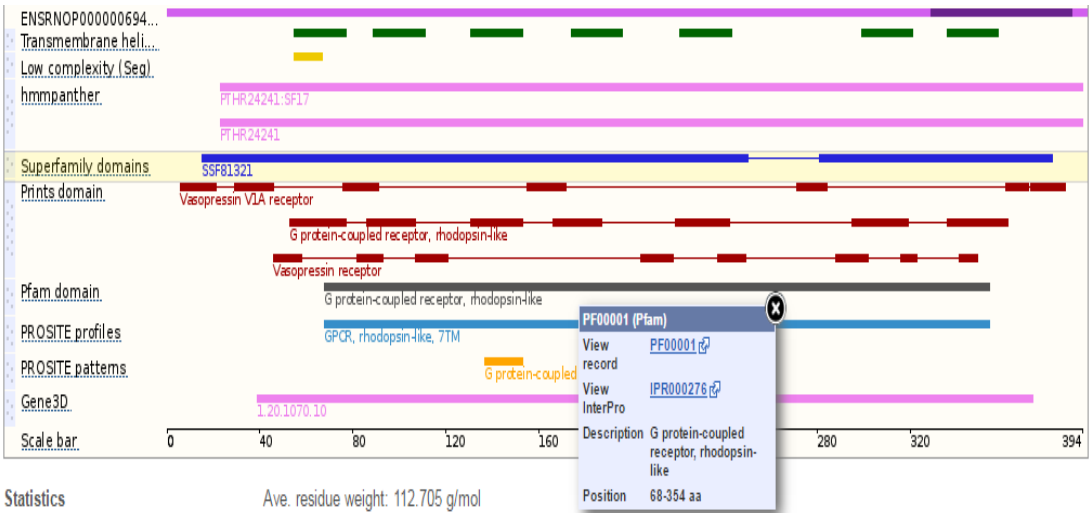


Figure 5.14: Summary for transcript: ENSRNOT00000087045, gene: Avpr1a (data retrieved in November, 2016)

Chapter 6

CONCLUSION

6.1 Main Findings

The main aim of this work was to find the amount of diversity that exists in the protein coded by the transcripts of GPCR genes in three different genomes, namely, human, mouse and rat. The study explored the availability of online platforms with comprehensive datasets about GPCR genes and proteins.

Our findings; based on the available data by November, 2016; indicate that GPCR proteome diversity is greatly influenced by the transcript diversity in all three genomes. Specifically, in human, 83% of GPCR genes with multiple transcripts code for variety of domains in their respective proteins. This percentage is about 81% for mouse and 65% for rat. These percentages are expected to increase with accumulating transcript data, as well as with expansion of protein domain databases.

In short, this work confirmed that there are variations in the domains coded by transcripts of most GPCR genes with multiple transcripts. This knowledge adds to the ever-growing field of GPCR drug target research, by confirming the gene expression complexity and proteome diversity of this group of proteins.

Biological databases are ever growing, and so is research on GPCRs. This work adds an important detail to GPCR research, potentially in the area of finding how and how

much drugs acts or bind to GPCRs; providing a direction to this research and other related work because the presence of differences in protein domains coded by transcripts gives rise to variation in structure and therefore functions.

6.2 Future Directions

The outcome of this thesis and works encountered during the course of this work has motivated some potential future directions: firstly, further study on factors leading to the domain diversity listed above will be considered because this might go a long way in helping in the development of drug that work by binding to GPCRs in the future. Secondly, further elucidation of differential expression of transcript diversity under normal or pathological conditions could be investigated. In addition, an analysis of mutations that change domain alterations could provide a potential biomedical direction. This is significant in drug target identification research. Furthermore, transcript diversity in GPCRs could be extended to other mammalian genomes.

It should also be noted that transcriptional diversity is under the influence of many factors; including developmental stage, tissue specificity, physiological condition and epigenetics. Therefore, more comprehensive approaches could be taken towards the study presented here and could involve analyses of how some or all of these factors play a role in expression of domain diversity.

REFERENCES

[1] Mandal A. What is Gene Expression. Retrieved from http://www.news-medical.net/life-sciences/What-is-Gene-Expression.aspx

[2] Lodish H., Berk A., Zipursky S. L. et al. (2000). *Molecular Cell Biology*. 4th edition, New York, WH: Freeman.

[3] Ensembl Genome Browser. (2016, July 16). Retrieved from http://www.ensembl.org/index.html

[4] The Central Dogma of Molecular Biology. (2016, December 13). *MMG 233 2014 Genetics & Genomics Wiki*. Retrieved from http://mmg-233-2014-genetics-genomics.wikia.com/wiki/The_Central_Dogma_of_Molecular_Biology

[5] Gonzalo S. Nuclear Architecture. Department of Biochemistry & Molecular Biology, Saint Louis University School of Medicine.

[6] Clancy S. (2008). DNA transcription. *Nature Education* 1(1), 41.

[7] Brown T., and Brown D. Transcription, Translation and Replication. available online on http://www.atdbio.com/content/14/Transcription-Translation-and-Replication

[8] Walsh D., and Mohr I. (2011). Viral subversion of the host protein synthesis machinery. Nature Reviews Microbiology 9, 860-875. doi:10.1038/nrmicro2655

[9] de Napoles M., et al. (2004). Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. *Developmental Cell*. 7 (5), 663–676. doi:10.1016/j.devcel.2004.10.005

[10] Mercer T.R., and Mattick J.S. (2013). Understanding the regulatory and transcriptional complexity of the genome through structure. Genome Research. 23 (7), 1081–1088. doi:10.1101/gr.156612.113

[11] Struhl K. (1999). Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*. 98(1), 1–4. doi:10.1016/s0092-8674(00)805991

[12] Sonenberg N., and Hinnebusch A. G. (2009). Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell*, 136(4), 731–745. doi.org/10.1016/j.cell.2009.01.042

[13] Chhabra N. (2015). Regulation of gene expression in eukaryotes. *Lecture note.* Department of Biochemistry, S.S.R. Medical College, Mauritius.

[14] Ralston A., and Shaw K. (2008). Gene expression regulates cell differentiation. Nature Education 1(1):127

[15] Isea R. (2015). The Present-Day Meaning of the Word Bioinformatics. *Global Journal of Advanced Research.* 2(1), 70-73.

[16] Hogeweg P. (2011). The Roots of Bioinformatics in Theoretical Biology. *PLoS Comput Biol* 7(3): e1002021. doi:10.1371/journal.pcbi.1002021

[17] Thampi S.M. (2009). Introduction to Bioinformatics. *ARXIV*. arXiv:0911.4230

[18] The National Center for Biotechnology Information NCBI. (2016, July 11). Retrieved from https://www.ncbi.nlm.nih.gov/

[19] PRINTS. (2017, March 7). Retrieved from http://130.88.97.239/PRINTS/index.php

[20] Fosalli R., Joloobi T.A. (2016). Computational Method for Sequence analyzing. *International Journal for Research in Computer Science*. 2(7)

[21] The Mouse Genome. Nature publishing group. (2016, February 21). Retrieved from https://www.nature.com/nature/mousegenome/timeline/index.html

[22] Kumar S. (2005). Bioinformatics web. Retrieved January 2017 from: http://www.geocities.com/bioinformaticsweb/

[23] NCBI. (2017, March 7). PubMed. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/

[24] Pfam. (2017, March 7). Retrieved from http://pfam.sanger.ac.uk/family/PF03474

[25] PROSITES. (2017, March 7). Retrieved from http://prosite.expasy.org/

[26] Concepts of Bioinformatics. (2017, March 7). I.C.A.R. Indian Agricultural Statistics Research Institute. Retrieved from http://www.iasri.res.in/ebook/CAFT_sd/Concepts%20of%20Bioinformatics.pdf

[27] Bacardit J. Introduction to Bioinformatics. Interdisciplinary Computing and Complex Systems (ICOS). University of Nottingham.

[28] O'Connor C.M., and Adams J.U. (2010). *Essentials of Cell Biology*. Cambridge, MA: NPG Education.

[29] Saygin Y., Sezerman U., and Cobanoglu, M. (2011). Classification of GPCRs Using Family Specific Motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8, 1495-1508. doi:10.1109/TCBB.2010.101

[30] Katritch V., Cherezov V., and Stevens R.C. (2013). Structure-Function of the G-protein-Coupled Receptor Superfamily. *Annu Rev Pharmacol Toxicol* 53(1), 531–556. doi:10.1146/annurevpharmtox-032112-135923

[31] Katritch V., Cherezov V., and Stevens R.C. (2012). Diversity and modularity of G protein-coupled receptor structures. 33(1), 17–27.

[32] Wheatley M., et al. (2012). Lifting the lid on GPCRs: the role of extracellular loops. *British Journal of Pharmacology*, 165(6), 1688-1703. DOI: 10.1111/j.1476-5381.2011.01629.x

[33] BioMart. (2016, July 16). Retrieved from http://www.ensembl.org/biomart/martview/bb9b745659f6b988a0420d821523a4fb

[34] Universal Protein Resource (UniProt). (2016, December 12). Retrieved from http://www.uniprot.org/

[35] NCBI Resource Coordinators (2012). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 41 (Database issue): D8–D20.

[36] Basic Local Alignment Search Tool. (2016, September 22). Retrieved from https://blast.ncbi.nlm.nih.gov/Blast.cgi#

[37] Introducing: Magic-BLAST. (2016, September 22). Retrieved from https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastNews

[38] Sayers E. A (2010). General Introduction to the E-utilities. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK25497/#chapter2

[39] Entrez Programming Utilities Help. (2010). National Center for Biotechnology Information. Bethesda, MD.

[40] Spudich G.M., and Fernández-Suárez, X.M. (2010). Touring Ensembl: A practical guide to genome browsing. BMC Bioinformatics, 11:295.

[41] Extracting data with BioMart. (2016, April 30). Retrieved from available on http://www.ensembl.org/info/data/biomart/index.html

[42] Bringing MySQL to the web. (2017, March 2). Retrieved from https://www.phpmyadmin.net/about/

[43] Technosip Review – Xampp. Software Review. (2016, December 16). Retrieved from http://www.technosip.com/toolsreviews/technosip-review-xampp/

[44] The Perl for MS Windows. (2017, March 2). Retrieved from http://strawberryperl.com/

[45] Finn R.D., et al. (2014). Pfam: the protein families database. Nucleic Acids Research, D222–D230, Vol. 42, November 2013

[46] Canese K., Jentsch J., and Myers J.C. (2002). The NCBI handbook 2nd edition, Chapter 2. PubMed. The Bibliographic Database.

[47] Beck J., and Sequeira E. (2002). *The NCBI handbook* 2nd edition, Chapter 9, PubMed Central (PMC). An Archive for Literature from Life Sciences Journals

[48] Pundir S., Magrane M., Martin M.J., and O'Donovan C. (2015). Searching and Navigating UniProt Databases. *UniProt Consortium.* Curr. Protoc. Bioinformatics 50, 1.27.1-1.27.10

[49] Palmer M., Chan A., Dieckmann T., and Honek J. (2013). Notes to Biochemical Pharmacology, John Wiley & Sons

APPENDICES

Appendix A: Perl code for parsing file downloaded from UniProt

i. #print lines with $find

```
#and replaces /<accession>/ with space
use strict;
use warnings;

my $filename = 'C:\Users\babsf\Downloads\uniprotrat1.xml';
open(my $fh, '<:encoding(UTF-8)', $filename)
or die "Could not open file '$filename' $!";

my $find    = '<accession>';
my $replace = '';

while (my $line = <$fh>) {
   next unless $line =~ /<accession>/;
        $line =~ s/\Q$find\E/$replace/g;
        chomp $line;
   print "$line\n";
}
```

ii. #find </accession> and replace with space

```
use strict;
use warnings;

my $filename = 'C:\perl_tests\ura.txt';
open(my $fh, '<:encoding(UTF-8)', $filename)
or die "Could not open file '$filename' $!";

my $find    = '</accession>';
my $replace = '';
while (my $line = <$fh>) {
        chomp $line;  #remove empty lines
        $line =~ s/\Q$find\E/$replace/g;
        $line =~ s/\s//g;  #remove white space
        chomp $line;
   print "$line\n";
}
```

Appendix B: List of protein IDs of human GPCR family in UniProt

P30988	P21731	Q9NYQ7	P48546
Q8NFJ5	Q96PE1	P08588	P43115
Q9H1Y3	Q9Y5X5	Q9HBW0	Q16602
O95838	Q8IZF2	P21917	O94910
P21452	Q9NYQ6	P41586	O00398
P47872	P41145	O60431	P46663
Q9Y5N1	P30989	Q8NFJ6	P51810
P04000	Q13585	P32247	P28223
O43193	Q9NPG1	P41587	Q13324
Q02643	P30968	P32239	Q9NZD1
Q9ULW2	Q6QNK2	P35372	P43220
Q9BZJ6	Q86SQ4	Q9UGF5	Q9UGF6
Q9UGF7	P51684	O15303	P25021
P32241	P25103	Q969F8	P41143
O95490	P21554	Q15743	O95800
P49683	P34998	O00270	P35348
Q9NYW2	Q9NYW0	Q9NYW3	Q9NYW1
O60241	O43613	O00421	P41597
Q92847	P61073	Q9UHM6	P43088
P48960	Q9NYM4	P41968	Q14833
P58173	Q6U736	Q14330	P43116
P41146	P48146	O95977	P30550
O15529	O15552	O14842	P32248
P30518	Q9NYW6	Q9NYW5	Q9NYW4
Q9UHX3	O76100	O60412	O76099
Q96RI0	P47887	P29274	Q99705
O60755	Q9NYV7	Q99835	P03999
Q9NPB9	P04201	Q8NG94	Q9BY15
P31391	Q9GZQ4	P47775	P20309
Q9HBX8	Q8WXD0	Q9H3N8	P49286
P35414	Q9P296	O60242	P28222
O14804	P28336	Q9Y2T6	P41595
Q9Y5Y3	P24530	Q8NGS3	Q8NGT0
O75899	Q8NGR5	O43614	P22888
Q9BY21	O75473	Q9NQS5	P43657
P30872	Q8IYL9	Q15615	O75388
Q9HCU4	Q8N6U8	Q13304	P49190
P25106	O00590	Q96K78	P30556
P46095	Q9P1P5	Q96RI8	Q969N4
Q96RJ0	Q9BXC1	Q99677	P28335
Q8TDV5	Q8NGS5	Q8NGR4	P47804
P34969	Q9GZQ6	O95221	P14416
Q8NGG5	Q8NGP3	Q9HAR2	P08913
Q5T848	P35462	P25101	Q969V1
Q9NS75	Q13255	Q8IWK6	Q8NGZ2
Q5T601	Q8IZF3	P49685	Q8NDV2
Q8NGD4	Q96P69	O75084	Q9BZJ8
Q8IZF6	P47898	Q9UP38	Q6NV75
P50406	Q8IZF4	P47888	P43119
P32302	P51681	Q03431	P34995
Q8NGA0	P35346	Q9HBW9	A3KFT3
Q8TDS5	Q13467	P32238	P25024
P30542	P32246	P08100	Q14416

P25929	Q15761	Q8WXG9	O00254
P55085	Q13639	Q8IZF7	P60893
Q96CH1	Q99527	O60353	Q15612
Q8NGR3	Q96P66	P16473	Q96P68
Q8NGD2	Q14439	P37288	P0DN77
Q9H207	Q9H210	P47211	P32245
O15218	Q8NGJ1	Q9NZH0	Q9H255
Q9H343	Q8NH59	Q8NGG0	Q13606
O43749	Q9GZK3	O95371	Q13258
P49238	P30411	P48039	P11229
Q8NGT2	P28566	P41594	O95665
Q8NGC4	Q8NGZ6	P07550	Q9H244
Q8NGC0	P25105	Q8NGD5	Q8NH41
P32249	Q8TCW9	P21918	Q9NYW7
P29371	P47900	Q7RTX0	O60883
O00155	P35368	Q96LB2	Q9NQ84
P29275	Q8NGE7	P58181	O15354
Q9H209	Q9H208	P23945	Q9NS67
Q8NG95	Q8NG99	Q8NGA1	Q8NGG6
P21453	Q8NGN0	P25090	P21462
Q8NG92	Q8NGT1	Q96R27	Q8NGR2
Q8NH93	Q8NGR8	Q8NGR9	Q8NGS0
Q9HBX9	Q9UBY5	P35408	A6NND4
Q9HB89	Q15077	Q96P67	Q9UPC5
Q16581	P25100	O60404	O60403
Q9H2C5	Q9P1Q5	Q9Y585	Q8N0Y5
Q6IEV9	Q8NH10	Q8NGF9	Q99680
O00574	Q8NGI7	Q8NGJ0	Q8NGI9
Q96R08	Q96R09	Q8NGQ6	Q8NGQ1
Q8NGP9	Q8NGF4	Q8NGP2	Q8NGG3
Q8NGA6	Q8NGE0	Q8NGE8	Q8NH48
Q8NGF6	Q96AM1	Q8TDS7	Q8TDU9
Q9Y271	Q8TDV0	Q8NGX5	Q8TDV2
Q8IZF5	P46094	P51686	Q5T6X5
Q7RTX1	Q8NH94	Q8IZP9	Q9UJ42
Q8NGV0	P34981	Q99788	Q8NG98
Q9ULV1	Q14246	Q8NH87	Q8NGP4
Q15391	O14626	Q86SP6	Q8NH18
Q8NH19	Q8NH73	O43869	Q8NGE3
Q6IF63	Q8TDT2	P41231	Q8NGF8
Q8NFN8	P33032	Q8NGC9	Q8NGI4
Q8NGC7	Q8NGC6	Q8NGA5	Q9H340
Q8NH43	Q8NH42	O95013	Q8NGD3
Q8NGD1	Q8NGD0	Q8NGI8	Q8NGB2
Q8NH37	Q8NGB4	Q8NH49	Q8NGA8
Q9H2C8	Q8NGJ4	Q8NGJ5	Q8NGK0
Q8NH64	Q8NGJ9	Q8NH63	Q8NGJ8
Q8NH61	Q8NGX2	Q8NHC7	Q8NG81
Q8NG77	Q8NG76	Q8NGZ0	Q9H461
A6NH00	P46093	Q8NGZ4	Q8NGZ5
Q5JQS5	Q96R69	Q8NGU9	Q9GZP7
Q8NGH3	P08908	Q8NGW1	Q9HC97
Q8TE23	Q8NGV5	P30939	Q8NGT5
P28221	O00222	Q8NGE2	Q8NGE1
Q9NZP2	Q96LA9	Q96LB0	Q8NGE5
Q8TDU6	P51685	P30953	P47881
O14718	Q7Z7M1	Q6DWJ6	Q14332
Q8NGW6	Q8NGQ4	P30874	O95047

Q6IF99	P08173	Q9H228	Q5UAW9
P50052	Q8TCB6	P25025	P30559
Q8NH76	Q9BZJ7	Q8NGH5	Q8NGH9
Q8NGI0	Q8NH53	Q8NH56	Q8NGI3
P08172	Q8NGI2	P25116	Q8NH90
Q8NGP6	Q9UKP6	Q8NGN1	Q8NGM9
Q9H346	Q8NGJ2	Q8IZ08	Q9BPV8
Q9GZN0	Q8NGG4	Q8NG75	Q8NGG2
Q8NG97	Q8NH50	Q8N146	Q8N162
P46089	Q8N127	O14514	Q8NH40
Q8NH72	Q8NGL7	Q8NGM1	Q8NGL6
Q8NH70	Q8NGK3	Q9UKL2	Q6IFH4
Q86VZ1	Q86SM5	Q8WZ94	Q96R30
Q9NSD7	Q8NGN6	Q8NH05	O43603

Appendix C: List of protein IDs of mouse GPCR family in UniProt

A2ARI4	A4FUQ5	B2RPY5	B7ZCC9
E9Q6I0	F8VQN3	O08530	O08675
O08707	O08786	O08790	O08858
O09047	O35161	O35214	O35457
O35599	O35659	O54689	O54798
O54799	O54897	O55040	O70342
O70421	O88319	O88410	O88416
O88495	O88536	O88537	O88634
O88721	O88853	O88854	O88855
P0C5I1	P12657	P15409	P18762
P21729	P21761	P23275	P25962
P28334	P29754	P29755	P30548
P30549	P30554	P30558	P30728
P30730	P30731	P30873	P30875
P30935	P30966	P30987	P30993
P32082	P32211	P32240	P32299
P32300	P32304	P33033	P33534
P33766	P34968	P34971	P34983
P34984	P35343	P35347	P35363
P35374	P35375	P35377	P35378
P35412	P35413	P41588	P41593
P42866	P43117	P43252	P47743
P47746	P47750	P47774	P47936
P47937	P48302	P49650	P49681
P51436	P51491	P51675	P51676
P51678	P51680	P51682	P51683
P52592	P55086	P56450	P56479
P56481	P56484	P56485	P56726
P58307	P58308	P58406	P59528
P59529	P59530	P59532	P60894
P61168	P61793	P70174	P70205
P70259	P70263	P70310	P83861
P97288	P97292	P97295	P97468
P97718	P97751	P97772	P97926
Q01337	Q01727	Q01776	Q02152
Q02284	Q04573	Q04683	Q0P543
Q0VAX9	Q0VDU3	Q14BI2	Q3KNA1
Q3SXG2	Q3U3F9	Q3U507	Q3U6B2
Q3UFD7	Q3UG50	Q3UG61	Q3UJF0
Q3UN16	Q3UVD5	Q3UVX5	Q3UVY1
Q3V3Z3	Q58Y75	Q5FWI2	Q5IXF8
Q5NCH9	Q5QD04	Q5QD05	Q5QD06
Q5QD07	Q5QD08	Q5QD09	Q5QD10
Q5QD11	Q5QD12	Q5QD13	Q5QD14
Q5QD15	Q5QD16	Q5QD17	Q60612
Q60613	Q60614	Q60748	Q60755
Q60878	Q60879	Q60881	Q60882
Q60883	Q60884	Q60885	Q60886
Q60887	Q60888	Q60889	Q60890
Q60891	Q60892	Q60894	Q60895
Q61041	Q61086	Q61088	Q61089
Q61089	Q61090	Q61091	Q61121
Q61125	Q61184	Q61212	Q61224

Q61549	Q61606	Q61614	Q61616
Q62035	Q62053	Q62463	Q64264
Q64326	Q68ED2	Q6F3F9	Q6IYF8
Q6NS65	Q6PI62	Q6R6I7	Q6TAC4
Q6VMN6	Q6VZZ7	Q6W049	Q6X632
Q6YNI2	Q71MR7	Q76JU9	Q7M707
Q7M709	Q7M710	Q7M711	Q7M712
Q7M713	Q7M715	Q7M718	Q7M720
Q7M721	Q7M722	Q7M723	Q7M724
Q7M725	Q7TMA4	Q7TN51	Q7TQA4
Q7TQA5	Q7TQA6	Q7TQB0	Q7TQB8
Q7TQB9	Q7TQN9	Q7TQP0	Q7TQP2
Q7TQP3	Q7TQP4	Q7TR96	Q7TRF3
Q7TT36	Q80SS6	Q80T41	Q80T62
Q80TR1	Q80TS3	Q80UC6	Q80UC8
Q80WT4	Q80ZF8	Q810W6	Q8BFQ3
Q8BFU7	Q8BG55	Q8BGE9	Q8BKG4
Q8BL07	Q8BLD9	Q8BLG2	Q8BM96
Q8BMC0	Q8BMP4	Q8BUD0	Q8BX79
Q8BXS7	Q8BXS7	Q8BYC4	Q8BZ39
Q8BZA7	Q8BZL4	Q8BZP8	Q8BZR0
Q8C010	Q8C206	Q8C419	Q8CGM1
Q8CIP3	Q8CJ12	Q8JZZ7	Q8K087
Q8K0Z9	Q8K1Z6	Q8K209	Q8K458
Q8K4Z6	Q8R0T6	Q8VBS7	Q8VBV9
Q8VBW9	Q8VCJ6	Q8VCK6	Q8VEC3
Q8VES2	Q8VEW5	Q8VEW6	Q8VEX5
Q8VEY3	Q8VEZ0	Q8VF12	Q8VF13
Q8VF65	Q8VF66	Q8VF76	Q8VFC9
Q8VFD0	Q8VFD1	Q8VFD2	Q8VFD3
Q8VFK1	Q8VFK2	Q8VFK7	Q8VFL5
Q8VFL9	Q8VFM9	Q8VFV4	Q8VFX2
Q8VG02	Q8VG03	Q8VG04	Q8VG05
Q8VG06	Q8VG07	Q8VG08	Q8VG09
Q8VG13	Q8VG42	Q8VG43	Q8VG44
Q8VGD6	Q8VGI1	Q8VGI4	Q8VGI5
Q8VGI6	Q8VGK5	Q8VGQ7	Q8VGR9
Q8VGS3	Q8VIC7	Q8VIC9	Q8VIH9
Q91V45	Q91V95	Q91WW2	Q91X56
Q91ZB5	Q91ZB7	Q91ZB8	Q91ZB9
Q91ZC0	Q91ZC1	Q91ZC6	Q91ZE5
Q91ZI0	Q91ZI0	Q91ZV8	Q91ZY2
Q91ZZ5	Q920A1	Q920H4	Q923X1
Q923Y8	Q923Z0	Q924H0	Q924I3
Q925D8	Q925I4	Q99JA4	Q99JG2
Q99LE2	Q99MT6	Q99MT7	Q99MT8
Q99P50	Q9CPV9	Q9D8I2	Q9EP51
Q9EP66	Q9EP79	Q9EPB7	Q9EPB8
Q9EQ16	Q9EQ31	Q9EQ45	Q9EQ46
Q9EQ47	Q9EQ48	Q9EQ52	Q9EQD0
Q9EQQ3	Q9EQQ4	Q9ERK9	Q9ERZ3
Q9ERZ4	Q9ES90	Q9ESG6	Q9JHB2
Q9JIL6	Q9JIP6	Q9JJL9	Q9JJS7
Q9JKA3	Q9JKL1	Q9JKT3	Q9JKT4
Q9JL21	Q9QXZ9	Q9QY00	Q9QY42
Q9QY96	Q9QYS2	Q9R1C8	Q9R1K6
Q9R1W5	Q9R216	Q9WU02	Q9WUF1
Q9WUK7	Q9WUT7	Q9WV08	Q9WV18

Q9Z0D9	Q9Z0L1	Q9Z0U9	Q9Z1P4
Q9Z1V0	Q9Z282	Q9Z2B3	Q9Z2J6

Appendix D: List of protein IDs of rat GPCR family in UniProt

P35349	Q70VB1	Q9Z0U4	Q8K418
P30543	P48442	Q9QYC6	P35365
P30680	O35811	Q9EQD2	Q9JII9
P25099	Q9Z0W0	P70597	D4AC13
P35000	P30560	P32244	P97583
Q7TQN7	P30937	P35353	P31422
Q9JKE9	Q9JKT5	Q9Z2H4	Q9JKT7
P70536	P14600	Q64017	P08911
P35364	P70585	P30936	Q56UD9
P34975	P20272	P61169	Q9JHG3
Q9QZN9	P22086	P16177	Q66H29
P43140	Q8R416	Q9Z0R8	P14842
P33533	P29089	Q9JJH3	P21451
P47866	Q9WVT0	P56719	Q02644
P28565	P32215	Q67ET0	A1A5S3
P26255	P28564	P31421	P46002
Q7TN41	P61794	P48303	P56718
P19328	Q9EPX4	Q9ESQ4	Q63634
P49651	P35407	P23385	P83858
Q924Y8	Q7TN38	P12526	Q924T9
Q8K5E0	Q4G072	P28647	P43219
Q923Y5	Q4KLH9	Q62805	P35345
P43253	P35370	P20395	Q9QZH0
Q5FVG1	Q9JKE7	Q63645	P33535
P25095	P08482	P35411	P70596
P32305	P23267	Q63371	Q923K1
O89039	Q924U0	Q9R0Q2	P34978
P47752	Q9JKM5	P25961	Q76EI6
Q67ER8	Q67ES5	Q67ET3	Q2AC31
Q9JKU0	P0C0W8	O08725	Q7TQN8
P60895	D4A7K7	Q09QM4	Q5QD25
Q5QD24	Q923Y6	Q80Z39	Q67ES3
Q9ESP4	Q7TSN5	Q67ES7	Q67ET2
Q67ER9	P20789	O88917	P23270
Q5J3M3	Q675B7	Q6GUG4	Q923Y7
Q67ET7	Q67ES1	Q67ES6	Q5J3M4
Q67ET1	P35898	Q67ES9	Q8CJ11
Q675B9	Q7TN45	Q675B8	Q9ESC1
Q6XKD3	P49684	P30082	B2GV46
Q924T8	Q6Y1R5	P30951	Q5QD23
Q5Y4N8	Q923X8	P43118	P10980
Q9QXI3	P97520	Q5QD21	P70526
Q923Y4	P28646	P70612	Q923Y1
Q5J3G9	Q695P6	P08483	P51651
P35896	P23266	P31388	P35351
P16610	P28647	Q62928	O35476
Q5J3E5	Q67ES3	O88278	Q8R456
Q63118	Q7TN40	Q5J3L7	P32305
Q5J3N1	Q5J3F6	Q80T02	Q5J3L4
Q498S8	Q00788	Q5J3M9	P19020
Q63447	Q498S8	O08726	P23270
Q9Z0R7			

Appendix E: Domain per transcript SQL procedure for mouse

```
"CREATE PROCEDURE `mouseDomain`(IN `num` INT UNSIGNED) NOT DETERMINISTIC NO SQL SQL SECURITY DEFINER
SELECT g.GeneID, t.TranscriptID, COUNT(DISTINCT d.PfamID) as DomainCount
FROM Mousegene as g, Mousetranscript as t, Mousedomain as d
WHERE g.GeneID = t.GeneID and t.TranscriptID = d.TranscriptID
GROUP BY g.GeneID, t.TranscriptID
HAVING DomainCount = num"
```

Where num is the variable for number of domains (from 1 to 8) as seen in the first column of table 4.7

Appendix F: Domain per transcript SQL procedure for rat

```
"CREATE PROCEDURE `ratDomain`(IN `num` INT UNSIGNED)
NOT DETERMINISTIC NO SQL SQL SECURITY DEFINER
SELECT g.GeneID, t.TranscriptID, COUNT(DISTINCT d.PfamID)
as DomainCount
FROM Ratgene as g, Rattranscript as t, Ratdomain as d
WHERE g.GeneID = t.GeneID and t.TranscriptID = d.TranscriptID
GROUP BY g.GeneID, t.TranscriptID
HAVING DomainCount = num"
```

Where num is the variable for number of domains (from 1 to 8) as seen in the first column of table 4.7

Appendix G: Human GPCRs with domain diversity

Gene ID	Description	Associated Gene Name
ENSG00000006638	thromboxane A2 receptor	TBXA2R
ENSG00000008300	cadherin EGF LAG seven-pass G-type receptor 3	CELSR3
ENSG00000010310	gastric inhibitory polypeptide receptor	GIPR
ENSG00000013588	G protein-coupled receptor class C group 5 member A	GPRC5A
ENSG00000020181	adhesion G protein-coupled receptor A2	ADGRA2
ENSG00000050628	prostaglandin E receptor 3	PTGER3
ENSG00000054277	opsin 3	OPN3
ENSG00000056291	neuropeptide FF receptor 2	NPFFR2
ENSG00000064547	lysophosphatidic acid receptor 2	LPAR2
ENSG00000064989	calcitonin receptor like receptor	CALCRL
ENSG00000065325	glucagon like peptide 2 receptor	GLP2R
ENSG00000069122	adhesion G protein-coupled receptor F5	ADGRF5
ENSG00000069696	dopamine receptor D4	DRD4
ENSG00000072071	adhesion G protein-coupled receptor L1	ADGRL1
ENSG00000075073	tachykinin receptor 2	TACR2
ENSG00000075275	cadherin EGF LAG seven-pass G-type receptor 1	CELSR1
ENSG00000078549	ADCYAP receptor type I	ADCYAP1R1
ENSG00000078589	purinergic receptor P2Y10	P2RY10
ENSG00000080293	secretin receptor	SCTR
ENSG00000082556	opioid receptor kappa 1	OPRK1
ENSG00000100739	bradykinin receptor B1	BDKRB1
ENSG00000101180	histamine receptor H3	HRH3
ENSG00000101188	neurotensin receptor 1	NTSR1
ENSG00000101850	G protein-coupled receptor 143	GPR143
ENSG00000102076	opsin 1 (cone pigments), long-wave-sensitive	OPN1LW
ENSG00000102468	5-hydroxytryptamine receptor 2A	HTR2A
ENSG00000104290	frizzled class receptor 3	FZD3
ENSG00000106018	vasoactive intestinal peptide receptor 2	VIPR2
ENSG00000106113	corticotropin releasing hormone receptor 2	CRHR2
ENSG00000106128	growth hormone releasing hormone receptor	GHRHR
ENSG00000110148	cholecystokinin B receptor	CCKBR
ENSG00000111291	G protein-coupled receptor class C group 5 member D	GPRC5D
ENSG00000111432	frizzled class receptor 10	FZD10
ENSG00000111452	adhesion G protein-coupled receptor D1	ADGRD1
ENSG00000112038	opioid receptor mu 1	OPRM1
ENSG00000112414	adhesion G protein-coupled receptor G6	ADGRG6
ENSG00000113262	glutamate metabotropic receptor 6	GRM6
ENSG00000113749	histamine receptor H2	HRH2
ENSG00000114812	vasoactive intestinal peptide receptor 1	VIPR1
ENSG00000115353	tachykinin receptor 1	TACR1
ENSG00000116014	KISS1 receptor	KISS1R
ENSG00000116329	opioid receptor delta 1	OPRD1

ENSG00000117114	adhesion G protein-coupled receptor L2	ADGRL2
ENSG00000118432	cannabinoid receptor 1	CNR1
ENSG00000119714	G protein-coupled receptor 68	GPR68
ENSG00000119973	prolactin releasing hormone receptor	PRLHR
ENSG00000120088	corticotropin releasing hormone receptor 1	CRHR1
ENSG00000120907	adrenoceptor alpha 1A	ADRA1A
ENSG00000121753	adhesion G protein-coupled receptor B2	ADGRB2
ENSG00000121764	hypocretin receptor 1	HCRTR1
ENSG00000121797	C-C motif chemokine receptor like 2	CCRL2
ENSG00000121807	C-C motif chemokine receptor 2	CCR2
ENSG00000121966	C-X-C motif chemokine receptor 4	CXCR4
ENSG00000122375	opsin 4	OPN4
ENSG00000122420	prostaglandin F receptor	PTGFR
ENSG00000123146	adhesion G protein-coupled receptor E5	ADGRE5
ENSG00000123901	G protein-coupled receptor 83	GPR83
ENSG00000124493	glutamate metabotropic receptor 4	GRM4
ENSG00000124818	opsin 5	OPN5
ENSG00000125245	G protein-coupled receptor 18	GPR18
ENSG00000125384	prostaglandin E receptor 2	PTGER2
ENSG00000125910	sphingosine-1-phosphate receptor 4	S1PR4
ENSG00000126262	free fatty acid receptor 2	FFAR2
ENSG00000126353	C-C motif chemokine receptor 7	CCR7
ENSG00000126895	arginine vasopressin receptor 2	AVPR2
ENSG00000127507	adhesion G protein-coupled receptor E2	ADGRE2
ENSG00000127533	F2R like thrombin/trypsin receptor 3	F2RL3
ENSG00000128271	adenosine A2a receptor	ADORA2A
ENSG00000128285	melanin concentrating hormone receptor 1	MCHR1
ENSG00000128602	smoothened, frizzled class receptor	SMO
ENSG00000129048	atypical chemokine receptor 4	ACKR4
ENSG00000131355	adhesion G protein-coupled receptor E3	ADGRE3
ENSG00000132911	neuromedin U receptor 2	NMUR2
ENSG00000133019	cholinergic receptor muscarinic 3	CHRM3
ENSG00000133067	leucine rich repeat containing G protein-coupled receptor 6	LGR6
ENSG00000134640	melatonin receptor 1B	MTNR1B
ENSG00000134817	apelin receptor	APLNR
ENSG00000135298	adhesion G protein-coupled receptor B3	ADGRB3
ENSG00000135577	neuromedin B receptor	NMBR
ENSG00000135898	G protein-coupled receptor 55	GPR55
ENSG00000136160	endothelin receptor type B	EDNRB
ENSG00000136928	gamma-aminobutyric acid type B receptor subunit 2	GABBR2
ENSG00000138039	luteinizing hormone/choriogonadotropin receptor	LHCGR
ENSG00000138271	G protein-coupled receptor 87	GPR87
ENSG00000139292	leucine rich repeat containing G protein-coupled receptor 5	LGR5
ENSG00000139679	lysophosphatidic acid receptor 6	LPAR6
ENSG00000143126	cadherin EGF LAG seven-pass G-type receptor 2	CELSR2
ENSG00000143147	G protein-coupled receptor 161	GPR161
ENSG00000144230	G protein-coupled receptor 17	GPR17
ENSG00000144407	parathyroid hormone 2 receptor	PTH2R

ENSG00000144476	atypical chemokine receptor 3	ACKR3
ENSG00000144648	atypical chemokine receptor 2	ACKR2
ENSG00000144820	adhesion G protein-coupled receptor G7	ADGRG7
ENSG00000144891	angiotensin II receptor type 1	AGTR1
ENSG00000147145	lysophosphatidic acid receptor 4	LPAR4
ENSG00000148604	retinal G protein coupled receptor	RGR
ENSG00000149295	dopamine receptor D2	DRD2
ENSG00000150471	adhesion G protein-coupled receptor L3	ADGRL3
ENSG00000151025	G protein-coupled receptor 158	GPR158
ENSG00000151577	dopamine receptor D3	DRD3
ENSG00000151617	endothelin receptor type A	EDNRA
ENSG00000152207	cysteinyl leukotriene receptor 2	CYSLTR2
ENSG00000152822	glutamate metabotropic receptor 1	GRM1
ENSG00000152990	adhesion G protein-coupled receptor A3	ADGRA3
ENSG00000153292	adhesion G protein-coupled receptor F1	ADGRF1
ENSG00000153294	adhesion G protein-coupled receptor F4	ADGRF4
ENSG00000155269	G protein-coupled receptor 78	GPR78
ENSG00000156097	G protein-coupled receptor 61	GPR61
ENSG00000157219	5-hydroxytryptamine receptor 5A	HTR5A
ENSG00000159618	adhesion G protein-coupled receptor G5	ADGRG5
ENSG00000160013	prostaglandin I2 (prostacyclin) receptor (IP)	PTGIR
ENSG00000160801	parathyroid hormone 1 receptor	PTH1R
ENSG00000162618	adhesion G protein-coupled receptor L4	ADGRL4
ENSG00000163485	adenosine A1 receptor	ADORA1
ENSG00000164082	glutamate metabotropic receptor 2	GRM2
ENSG00000164128	neuropeptide Y receptor Y1	NPY1R
ENSG00000164199	adhesion G protein-coupled receptor V1	ADGRV1
ENSG00000164251	F2R like trypsin receptor 1	F2RL1
ENSG00000164270	5-hydroxytryptamine receptor 4	HTR4
ENSG00000164393	adhesion G protein-coupled receptor F2	ADGRF2
ENSG00000164604	G protein-coupled receptor 85	GPR85
ENSG00000164849	G protein-coupled receptor 146	GPR146
ENSG00000164850	G protein-coupled estrogen receptor 1	GPER1
ENSG00000164930	frizzled class receptor 6	FZD6
ENSG00000165409	thyroid stimulating hormone receptor	TSHR
ENSG00000165621	oxoglutarate receptor 1	OXGR1
ENSG00000166073	G protein-coupled receptor 176	GPR176
ENSG00000166148	arginine vasopressin receptor 1A	AVPR1A
ENSG00000166160	opsin 1 (cone pigments), medium-wave-sensitive 2	OPN1MW2
ENSG00000166573	galanin receptor 1	GALR1
ENSG00000166856	G protein-coupled receptor 182	GPR182
ENSG00000167191	G protein-coupled receptor class C group 5 member B	GPRC5B
ENSG00000167332	olfactory receptor family 51 subfamily E member 2	OR51E2
ENSG00000168329	C-X3-C motif chemokine receptor 1	CX3CR1
ENSG00000168539	cholinergic receptor muscarinic 1	CHRM1
ENSG00000168959	glutamate metabotropic receptor 5	GRM5
ENSG00000169313	purinergic receptor P2Y12	P2RY12
ENSG00000169777	taste 2 receptor member 1	TAS2R1
ENSG00000170255	MAS related GPR family member X1	MRGPRX1
ENSG00000170412	G protein-coupled receptor class C group 5	GPRC5C

	member C	
ENSG00000170425	adenosine A2b receptor	ADORA2B
ENSG00000170820	follicle stimulating hormone receptor	FSHR
ENSG00000170989	sphingosine-1-phosphate receptor 1	S1PR1
ENSG00000171049	formyl peptide receptor 2	FPR2
ENSG00000171051	formyl peptide receptor 1	FPR1
ENSG00000171133	olfactory receptor family 2 subfamily K member 2	OR2K2
ENSG00000171459	olfactory receptor family 1 subfamily L member 6	OR1L6
ENSG00000171501	olfactory receptor family 1 subfamily N member 2	OR1N2
ENSG00000171505	olfactory receptor family 1 subfamily N member 1	OR1N1
ENSG00000171509	relaxin/insulin like family peptide receptor 1	RXFP1
ENSG00000171517	lysophosphatidic acid receptor 3	LPAR3
ENSG00000171522	prostaglandin E receptor 4	PTGER4
ENSG00000171631	pyrimidinergic receptor P2Y6	P2RY6
ENSG00000171657	G protein-coupled receptor 82	GPR82
ENSG00000171659	G protein-coupled receptor 34	GPR34
ENSG00000171860	complement C3a receptor 1	C3AR1
ENSG00000171873	adrenoceptor alpha 1D	ADRA1D
ENSG00000172209	G protein-coupled receptor 22	GPR22
ENSG00000172464	olfactory receptor family 5 subfamily AP member 2	OR5AP2
ENSG00000172935	MAS related GPR family member F	MRGPRF
ENSG00000173198	cysteinyl leukotriene receptor 1	CYSLTR1
ENSG00000173567	adhesion G protein-coupled receptor F3	ADGRF3
ENSG00000173578	X-C motif chemokine receptor 1	XCR1
ENSG00000173585	C-C motif chemokine receptor 9	CCR9
ENSG00000173662	taste 1 receptor member 1	TAS1R1
ENSG00000173679	olfactory receptor family 1 subfamily L member 1	OR1L1
ENSG00000173698	adhesion G protein-coupled receptor G2	ADGRG2
ENSG00000174600	chemerin chemokine-like receptor 1	CMKLR1
ENSG00000174837	adhesion G protein-coupled receptor E1	ADGRE1
ENSG00000174944	purinergic receptor P2Y14	P2RY14
ENSG00000174946	G protein-coupled receptor 171	GPR171
ENSG00000175697	G protein-coupled receptor 156	GPR156
ENSG00000176136	melanocortin 5 receptor	MC5R
ENSG00000176294	olfactory receptor family 4 subfamily N member 2	OR4N2
ENSG00000176547	olfactory receptor family 4 subfamily C member 3	OR4C3
ENSG00000176695	olfactory receptor family 4 subfamily F member 17	OR4F17
ENSG00000176900	olfactory receptor family 51 subfamily T member 1	OR51T1
ENSG00000177283	frizzled class receptor 8	FZD8
ENSG00000177464	G protein-coupled receptor 4	GPR4
ENSG00000178394	5-hydroxytryptamine receptor 1A	HTR1A
ENSG00000179603	glutamate metabotropic receptor 8	GRM8
ENSG00000179817	MAS related GPR family member X4	MRGPRX4

ENSG00000179826	MAS related GPR family member X3	MRGPRX3
ENSG00000179934	C-C motif chemokine receptor 8	CCR8
ENSG00000180090	olfactory receptor family 3 subfamily A member 1	OR3A1
ENSG00000180264	adhesion G protein-coupled receptor D2	ADGRD2
ENSG00000180269	G protein-coupled receptor 139	GPR139
ENSG00000180616	somatostatin receptor 2	SSTR2
ENSG00000180739	sphingosine-1-phosphate receptor 5	S1PR5
ENSG00000180758	G protein-coupled receptor 157	GPR157
ENSG00000180785	olfactory receptor family 51 subfamily E member 1	OR51E1
ENSG00000180871	C-X-C motif chemokine receptor 2	CXCR2
ENSG00000180914	oxytocin receptor	OXTR
ENSG00000181072	cholinergic receptor muscarinic 2	CHRM2
ENSG00000181104	coagulation factor II thrombin receptor	F2R
ENSG00000181619	G protein-coupled receptor 135	GPR135
ENSG00000181693	olfactory receptor family 8 subfamily H member 1	OR8H1
ENSG00000181767	olfactory receptor family 8 subfamily H member 2	OR8H2
ENSG00000181790	adhesion G protein-coupled receptor B1	ADGRB1
ENSG00000182162	purinergic receptor P2Y8	P2RY8
ENSG00000182854	olfactory receptor family 4 subfamily F member 15	OR4F15
ENSG00000182885	adhesion G protein-coupled receptor G3	ADGRG3
ENSG00000183150	G protein-coupled receptor 19	GPR19
ENSG00000183484	G protein-coupled receptor 132	GPR132
ENSG00000183625	C-C motif chemokine receptor 3	CCR3
ENSG00000183671	G protein-coupled receptor 1	GPR1
ENSG00000183840	G protein-coupled receptor 39	GPR39
ENSG00000184160	adrenoceptor alpha 2C	ADRA2C
ENSG00000184194	G protein-coupled receptor 173	GPR173
ENSG00000184451	C-C motif chemokine receptor 10	CCR10
ENSG00000184574	lysophosphatidic acid receptor 5	LPAR5
ENSG00000184984	cholinergic receptor muscarinic 5	CHRM5
ENSG00000185231	melanocortin 2 receptor	MC2R
ENSG00000186188	free fatty acid receptor 4	FFAR4
ENSG00000186867	pyroglutamylated RFamide peptide receptor	QRFPR
ENSG00000187037	G protein-coupled receptor 141	GPR141
ENSG00000187258	neuropeptide S receptor 1	NPSR1
ENSG00000188269	olfactory receptor family 7 subfamily A member 5	OR7A5
ENSG00000188691	olfactory receptor family 56 subfamily A member 5	OR56A5
ENSG00000188778	adrenoceptor beta 3	ADRB3
ENSG00000196131	vomeronasal 1 receptor 2	VN1R2
ENSG00000196277	glutamate metabotropic receptor 7	GRM7
ENSG00000196639	histamine receptor H1	HRH1
ENSG00000197376	olfactory receptor family 8 subfamily S member 1	OR8S1
ENSG00000197405	complement C5a receptor 1	C5AR1
ENSG00000198049	arginine vasopressin receptor 1B	AVPR1B
ENSG00000198121	lysophosphatidic acid receptor 1	LPAR1

ENSG00000198822	glutamate metabotropic receptor 3	GRM3
ENSG00000203757	olfactory receptor family 6 subfamily K member 3	OR6K3
ENSG00000204681	gamma-aminobutyric acid type B receptor subunit 1	GABBR1
ENSG00000204688	olfactory receptor family 2 subfamily H member 1	OR2H1
ENSG00000204689	olfactory receptor family 10 subfamily C member 1 (gene/pseudogene)	OR10C1
ENSG00000205213	leucine rich repeat containing G protein-coupled receptor 4	LGR4
ENSG00000205336	adhesion G protein-coupled receptor G1	ADGRG1
ENSG00000205409	olfactory receptor family 52 subfamily E member 6	OR52E6
ENSG00000206466	gamma-aminobutyric acid type B receptor subunit 1	GABBR1
ENSG00000206471	olfactory receptor family 2 subfamily H member 1	OR2H1
ENSG00000206474	olfactory receptor family 10 subfamily C member 1 (gene/pseudogene)	OR10C1
ENSG00000206511	gamma-aminobutyric acid type B receptor subunit 1	GABBR1
ENSG00000206516	olfactory receptor family 2 subfamily H member 1	OR2H1
ENSG00000212127	taste 2 receptor member 14	TAS2R14
ENSG00000213903	leukotriene B4 receptor	LTB4R
ENSG00000213906	leukotriene B4 receptor 2	LTB4R2
ENSG00000215644	glucagon receptor	GCGR
ENSG00000221938	olfactory receptor family 2 subfamily A member 14	OR2A14
ENSG00000224234	olfactory receptor family 10 subfamily C member 1 (gene/pseudogene)	OR10C1
ENSG00000224395	olfactory receptor family 2 subfamily H member 1	OR2H1
ENSG00000229125	olfactory receptor family 2 subfamily H member 1	OR2H1
ENSG00000229408	olfactory receptor family 2 subfamily H member 1	OR2H1
ENSG00000229412	olfactory receptor family 10 subfamily C member 1 (gene/pseudogene)	OR10C1
ENSG00000230505	olfactory receptor family 10 subfamily C member 1 (gene/pseudogene)	OR10C1
ENSG00000232268	olfactory receptor family 52 subfamily I member 1	OR52I1
ENSG00000232397	olfactory receptor family 10 subfamily C member 1 (gene/pseudogene)	OR10C1
ENSG00000232569	gamma-aminobutyric acid type B receptor subunit 1	GABBR1
ENSG00000232632	gamma-aminobutyric acid type B receptor subunit 1	GABBR1
ENSG00000232984	olfactory receptor family 2 subfamily H member 1	OR2H1
ENSG00000233412	olfactory receptor family 5 subfamily H member 15	OR5H15

ENSG00000235132	olfactory receptor family 2 subfamily H member 1	OR2H1
ENSG00000237051	gamma-aminobutyric acid type B receptor subunit 1	GABBR1
ENSG00000237112	gamma-aminobutyric acid type B receptor subunit 1	GABBR1
ENSG00000244165	purinergic receptor P2Y11	P2RY11
ENSG00000250510	G protein-coupled receptor 162	GPR162
ENSG00000257008	G protein-coupled receptor 142	GPR142
ENSG00000258839	melanocortin 1 receptor	MC1R
ENSG00000261984	taste 2 receptor member 14	TAS2R14
ENSG00000262111	taste 2 receptor member 30	TAS2R30
ENSG00000262611	olfactory receptor family 8 subfamily H member 1	OR8H1
ENSG00000263097	taste 2 receptor member 31	TAS2R31
ENSG00000268221	opsin 1 (cone pigments), medium-wave-sensitive	OPN1MW
ENSG00000276191	corticotropin releasing hormone receptor 1	CRHR1
ENSG00000276541	taste 2 receptor member 14	TAS2R14
ENSG00000278232	corticotropin releasing hormone receptor 1	CRHR1
ENSG00000280021	olfactory receptor family 51 subfamily F member 1 (gene/pseudogene)	OR51F1
ENSG00000282608	adenosine A3 receptor	ADORA3

Appendix H: Mouse GPCRs with domain diversity

Gene ID	Description	Associated Gene Name
ENSMUSG00000000617	glutamate receptor, metabotropic 6	Grm6
ENSMUSG00000000766	opioid receptor, mu 1	Oprm1
ENSMUSG00000001761	smoothened, frizzled class receptor	Smo
ENSMUSG00000003476	corticotropin releasing hormone receptor 2	Crhr2
ENSMUSG00000004654	growth hormone releasing hormone receptor	Ghrhr
ENSMUSG00000006378	glycine C-acetyltransferase (2-amino-3-ketobutyrate-coenzyme A ligase)	Gcat
ENSMUSG00000007989	frizzled class receptor 3	Fzd3
ENSMUSG00000008734	G protein-coupled receptor, family C, group 5, member B	Gprc5b
ENSMUSG00000011171	vasoactive intestinal peptide receptor 2	Vipr2
ENSMUSG00000013033	adhesion G protein-coupled receptor L1	Adgrl1
ENSMUSG00000019464	prostaglandin E receptor 1 (subtype EP1)	Ptger1
ENSMUSG00000019828	glutamate receptor, metabotropic 1	Grm1
ENSMUSG00000019865	neuromedin B receptor	Nmbr
ENSMUSG00000020140	leucine rich repeat containing G protein coupled receptor 5	Lgr5
ENSMUSG00000020793	galanin receptor 2	Galr2
ENSMUSG00000020963	thyroid stimulating hormone receptor	Tshr
ENSMUSG00000021303	guanine nucleotide binding protein (G protein), gamma 4	Gng4
ENSMUSG00000023192	glutamate receptor, metabotropic 2	Grm2
ENSMUSG00000023439	guanine nucleotide binding protein (G protein), beta 3	Gnb3
ENSMUSG00000023473	cadherin, EGF LAG seven-pass G-type receptor 3	Celsr3
ENSMUSG00000024211	glutamate receptor, metabotropic 8	Grm8
ENSMUSG00000024462	gamma-aminobutyric acid (GABA) B receptor, 1	Gabbr1
ENSMUSG00000024798	5-hydroxytryptamine (serotonin) receptor 7	Htr7
ENSMUSG00000025127	glucagon receptor	Gcgr
ENSMUSG00000025333	G protein-coupled receptor 143	Gpr143
ENSMUSG00000025475	adhesion G protein-coupled receptor A1	Adgra1
ENSMUSG00000025496	dopamine receptor D4	Drd4
ENSMUSG00000025905	opioid receptor, kappa 1	Oprk1
ENSMUSG00000025946	parathyroid hormone 2 receptor	Pth2r
ENSMUSG00000026228	5-hydroxytryptamine (serotonin) receptor 2B	Htr2b
ENSMUSG00000026237	neuromedin U receptor 1	Nmur1
ENSMUSG00000026271	G protein-coupled receptor 35	Gpr35
ENSMUSG00000026387	secretin receptor	Sctr
ENSMUSG00000026432	arginine vasopressin receptor 1B	Avpr1b
ENSMUSG00000027584	opioid receptor-like 1	Oprl1
ENSMUSG00000027669	guanine nucleotide binding protein (G protein), beta 4	Gnb4
ENSMUSG00000027762	succinate receptor 1	Sucnr1
ENSMUSG00000028004	neuropeptide Y receptor Y2	Npy2r
ENSMUSG00000028012	retinal pigment epithelium derived rhodopsin	Rrh

	homolog	
ENSMUSG00000028036	prostaglandin F receptor	Ptgfr
ENSMUSG00000028184	adhesion G protein-coupled receptor L2	Adgrl2
ENSMUSG00000028738	taste receptor, type 1, member 2	Tas1r2
ENSMUSG00000028782	adhesion G protein-coupled receptor B2	Adgrb2
ENSMUSG00000029064	guanine nucleotide binding protein (G protein), beta 1	Gnb1
ENSMUSG00000029090	adhesion G protein-coupled receptor A3	Adgra3
ENSMUSG00000029193	cholecystokinin A receptor	Cckar
ENSMUSG00000029255	gonadotropin releasing hormone receptor	Gnrhr
ENSMUSG00000029530	chemokine (C-C motif) receptor 9	Ccr9
ENSMUSG00000029663	guanine nucleotide binding protein (G protein), gamma transducing activity polypeptide 1	Gngt1
ENSMUSG00000029713	guanine nucleotide binding protein (G protein), beta 2	Gnb2
ENSMUSG00000029778	adenylate cyclase activating polypeptide 1 receptor 1	Adcyap1r1
ENSMUSG00000030043	tachykinin receptor 1	Tacr1
ENSMUSG00000030324	rhodopsin	Rho
ENSMUSG00000030406	gastric inhibitory polypeptide receptor	Gipr
ENSMUSG00000030898	cholecystokinin B receptor	Cckbr
ENSMUSG00000031070	MAS-related GPR, member F	Mrgprf
ENSMUSG00000031298	adhesion G protein-coupled receptor G2	Adgrg2
ENSMUSG00000031390	arginine vasopressin receptor 2	Avpr2
ENSMUSG00000031394	opsin 1 (cone pigments), medium-wave-sensitive (color blindness, deutan)	Opn1mw
ENSMUSG00000031486	adhesion G protein-coupled receptor A2	Adgra2
ENSMUSG00000031489	adrenergic receptor, beta 3	Adrb3
ENSMUSG00000031616	endothelin receptor type A	Ednra
ENSMUSG00000031785	adhesion G protein-coupled receptor G1	Adgrg1
ENSMUSG00000031932	G protein-coupled receptor 83	Gpr83
ENSMUSG00000032360	hypocretin (orexin) receptor 2	Hcrtr2
ENSMUSG00000032492	parathyroid hormone 1 receptor	Pth1r
ENSMUSG00000032528	vasoactive intestinal peptide receptor 1	Vipr1
ENSMUSG00000032641	G protein-coupled receptor 19	Gpr19
ENSMUSG00000032773	cholinergic receptor, muscarinic 1, CNS	Chrm1
ENSMUSG00000033569	adhesion G protein-coupled receptor B3	Adgrb3
ENSMUSG00000034009	relaxin/insulin-like family peptide receptor 1	Rxfp1
ENSMUSG00000034677	G protein-coupled receptor 142	Gpr142
ENSMUSG00000034730	adhesion G protein-coupled receptor B1	Adgrb1
ENSMUSG00000034987	histamine receptor H2	Hrh2
ENSMUSG00000036353	purinergic receptor P2Y, G-protein coupled 12	P2ry12
ENSMUSG00000036381	purinergic receptor P2Y, G-protein coupled, 14	P2ry14
ENSMUSG00000036402	guanine nucleotide binding protein (G protein), gamma 12	Gng12
ENSMUSG00000036437	neuropeptide Y receptor Y1	Npy1r
ENSMUSG00000037605	adhesion G protein-coupled receptor L3	Adgrl3
ENSMUSG00000038390	G protein-coupled receptor 162	Gpr162
ENSMUSG00000038607	guanine nucleotide binding protein (G protein), gamma 10	Gng10

ENSMUSG00000038668	lysophosphatidic acid receptor 1	Lpar1
ENSMUSG00000039059	histamine receptor H3	Hrh3
ENSMUSG00000039116	adhesion G protein-coupled receptor G6	Adgrg6
ENSMUSG00000039167	adhesion G protein-coupled receptor L4	Adgrl4
ENSMUSG00000039809	gamma-aminobutyric acid (GABA) B receptor, 2	Gabbr2
ENSMUSG00000039942	prostaglandin E receptor 4 (subtype EP4)	Ptger4
ENSMUSG00000040328	olfactory receptor 56	Olfr56
ENSMUSG00000040372	G protein-coupled receptor 63	Gpr63
ENSMUSG00000040836	G protein-coupled receptor 161	Gpr161
ENSMUSG00000041347	bradykinin receptor, beta 1	Bdkrb1
ENSMUSG00000041380	5-hydroxytryptamine (serotonin) receptor 2C	Htr2c
ENSMUSG00000041468	G-protein coupled receptor 12	Gpr12
ENSMUSG00000041907	G protein-coupled receptor 45	Gpr45
ENSMUSG00000042190	chemokine-like receptor 1	Cmklr1
ENSMUSG00000042429	adenosine A1 receptor	Adora1
ENSMUSG00000042793	leucine-rich repeat-containing G protein-coupled receptor 6	Lgr6
ENSMUSG00000042804	G protein-coupled receptor 153	Gpr153
ENSMUSG00000043017	prostaglandin I receptor (IP)	Ptgir
ENSMUSG00000043366	olfactory receptor 78	Olfr78
ENSMUSG00000043441	G protein-coupled receptor 149	Gpr149
ENSMUSG00000043659	neuropeptide S receptor 1	Npsr1
ENSMUSG00000043953	chemokine (C-C motif) receptor-like 2	Ccrl2
ENSMUSG00000044014	neuropeptide Y receptor Y5	Npy5r
ENSMUSG00000044017	adhesion G protein-coupled receptor D1	Adgrd1
ENSMUSG00000044067	G protein-coupled receptor 22	Gpr22
ENSMUSG00000044197	G protein-coupled receptor 146	Gpr146
ENSMUSG00000044288	cannabinoid receptor 1 (brain)	Cnr1
ENSMUSG00000044337	atypical chemokine receptor 3	Ackr3
ENSMUSG00000044338	apelin receptor	Aplnr
ENSMUSG00000044454	olfactory receptor 867	Olfr867
ENSMUSG00000045613	cholinergic receptor, muscarinic 2, cardiac	Chrm2
ENSMUSG00000045875	adrenergic receptor, alpha 1a	Adra1a
ENSMUSG00000045967	G protein-coupled receptor 158	Gpr158
ENSMUSG00000046159	cholinergic receptor, muscarinic 3, cardiac	Chrm3
ENSMUSG00000046793	G protein-coupled receptor 61	Gpr61
ENSMUSG00000046856	G protein-coupled receptor 1	Gpr1
ENSMUSG00000047415	G protein-coupled receptor 68	Gpr68
ENSMUSG00000047444	olfactory receptor 139	Olfr139
ENSMUSG00000047960	olfactory receptor 186	Olfr186
ENSMUSG00000048101	olfactory receptor 19	Olfr19
ENSMUSG00000048216	G protein-coupled receptor 85	Gpr85
ENSMUSG00000048240	guanine nucleotide binding protein (G protein), gamma 7	Gng7
ENSMUSG00000048779	pyrimidinergic receptor P2Y, G-protein coupled, 6	P2ry6
ENSMUSG00000049112	oxytocin receptor	Oxtr
ENSMUSG00000049115	angiotensin II receptor, type 1a	Agtr1a
ENSMUSG00000049130	complement component 5a receptor 1	C5ar1
ENSMUSG00000049409	prokineticin receptor 1	Prokr1
ENSMUSG00000049583	glutamate receptor, metabotropic 5	Grm5
ENSMUSG00000049649	G-protein coupled receptor 3	Gpr3

ENSMUSG00000049928	glucagon-like peptide 2 receptor	Glp2r
ENSMUSG00000050147	coagulation factor II (thrombin) receptor-like 3	F2rl3
ENSMUSG00000050199	leucine-rich repeat-containing G protein-coupled receptor 4	Lgr4
ENSMUSG00000050558	prokineticin receptor 2	Prokr2
ENSMUSG00000050921	purinergic receptor P2Y, G-protein coupled 10	P2ry10
ENSMUSG00000051314	free fatty acid receptor 2	Ffar2
ENSMUSG00000051431	G protein-coupled receptor 87	Gpr87
ENSMUSG00000051980	calcium-sensing receptor	Casr
ENSMUSG00000052270	formyl peptide receptor 2	Fpr2
ENSMUSG00000052303	MAS-related GPR, member A6	Mrgpra6
ENSMUSG00000053004	histamine receptor H1	Hrh1
ENSMUSG00000053164	G protein-coupled receptor 21	Gpr21
ENSMUSG00000053368	relaxin/insulin-like family peptide receptor 2	Rxfp2
ENSMUSG00000053647	G protein-coupled estrogen receptor 1	Gper1
ENSMUSG00000053852	adhesion G protein-coupled receptor G4	Adgrg4
ENSMUSG00000054141	olfactory receptor 24	Olfr24
ENSMUSG00000054764	melatonin receptor 1A	Mtnr1a
ENSMUSG00000056380	G-protein-coupled receptor 50	Gpr50
ENSMUSG00000056679	G-protein coupled receptor 173	Gpr173
ENSMUSG00000056755	glutamate receptor, metabotropic 7	Grm7
ENSMUSG00000058400	pyroglutamylated RFamide peptide receptor	Qrfpr
ENSMUSG00000058831	opsin 1 (cone pigments), short-wave-sensitive (color blindness, tritan)	Opn1sw
ENSMUSG00000059588	calcitonin receptor-like	Calcrl
ENSMUSG00000060470	adhesion G protein-coupled receptor G3	Adgrg3
ENSMUSG00000061577	adhesion G protein-coupled receptor G5	Adgrg5
ENSMUSG00000063120	olfactory receptor 480	Olfr480
ENSMUSG00000063594	guanine nucleotide binding protein (G protein), gamma 8	Gng8
ENSMUSG00000066896	olfactory receptor 18	Olfr18
ENSMUSG00000067642	adhesion G protein-coupled receptor F3	Adgrf3
ENSMUSG00000068037	MAS1 oncogene	Mas1
ENSMUSG00000068122	angiotensin II receptor, type 2	Agtr2
ENSMUSG00000068234	vomeronasal 1 receptor 44	Vmn1r44
ENSMUSG00000068523	guanine nucleotide binding protein (G protein), gamma 5	Gng5
ENSMUSG00000068696	G-protein coupled receptor 88	Gpr88
ENSMUSG00000069823	olfactory receptor 1	Olfr1
ENSMUSG00000070547	MAS-related GPR, member B1	Mrgprb1
ENSMUSG00000070687	5-hydroxytryptamine (serotonin) receptor 1D	Htr1d
ENSMUSG00000073008	G protein-coupled receptor 174	Gpr174
ENSMUSG00000074109	MAS-related GPR, member X2	Mrgprx2
ENSMUSG00000078118	olfactory receptor 483	Olfr483
ENSMUSG00000079227	chemokine (C-C motif) receptor 5	Ccr5
ENSMUSG00000079355	atypical chemokine receptor 4	Ackr4
ENSMUSG00000090951	olfactory receptor 181	Olfr181
ENSMUSG00000094426	olfactory receptor 478	Olfr478
ENSMUSG00000094612	olfactory receptor 491	Olfr491
ENSMUSG00000095212	olfactory receptor 473	Olfr473

ENSMUSG00000095239	olfactory receptor 497	Olfr497
ENSMUSG00000096209	olfactory receptor 510	Olfr510
ENSMUSG00000096465	olfactory receptor 488	Olfr488
ENSMUSG00000107269	cadherin, EGF LAG seven-pass G-type receptor 3	Celsr3

Appendix I: Rat GPCRs with domain diversity

Gene ID	Description	Associated Gene Name
ENSRNOG00000000774	gamma-aminobutyric acid type B receptor subunit 1	Gabbr1
ENSRNOG00000003305	C-X-C motif chemokine receptor 3	Cxcr3
ENSRNOG00000004400	arginine vasopressin receptor 1A	Avpr1a
ENSRNOG00000004900	corticotropin releasing hormone receptor 1	Crhr1
ENSRNOG00000005519	glutamate metabotropic receptor 3	Grm3
ENSRNOG00000011145	corticotropin releasing hormone receptor 2	Crhr2
ENSRNOG00000011154	adhesion G protein-coupled receptor F5	Adgrf5
ENSRNOG00000011808	growth hormone releasing hormone receptor	Ghrhr
ENSRNOG00000014269	C-X-C motif chemokine receptor 2	Cxcr2
ENSRNOG00000014290	glutamate metabotropic receptor 1	Grm1
ENSRNOG00000016768	opioid related nociceptin receptor 1	Oprl1
ENSRNOG00000018191	opioid receptor, mu 1	Oprm1
ENSRNOG00000018827	5-hydroxytryptamine receptor 7	Htr7
ENSRNOG00000029134	adhesion G protein-coupled receptor L1	Adgrl1
ENSRNOG00000036692	glucagon receptor	Gcgr
ENSRNOG00000037845	cysteinyl leukotriene receptor 1	Cysltr1
ENSRNOG00000049761	5-hydroxytryptamine receptor 6	Htr6
ENSRNOG00000050743	adenosine receptor A3-like	LOC100911796
ENSRNOG00000053893	opsin 4	Opn4
ENSRNOG00000055705	5-hydroxytryptamine receptor 7	LOC103694905
ENSRNOG00000059862	arginine vasopressin receptor 2	Avpr2
ENSRNOG00000061876	taste 1 receptor member 2	Tas1r2