A Computational Analysis of the Impact of Transcript Diversity on Protein Domains Coded by Human, Mouse and Rat Transcription Factor Genes

Salma Samiei

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
February 2014
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Elvan Yılmaz
Director

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

Prof. Dr. 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. Assoc. Prof. Dr. Bahar Taneri ____________________

2. Assoc. Prof. Dr. Ekrem Varoğlu ____________________

3. Asst. Prof. Dr. Nazife Dimililer ____________________

4. Asst. Prof. Dr. Mevhibe B. Hocaoğlu ____________________

5. Asst. Prof. Dr. Önsen Toygar ____________________

ABSTRACT

In this study, three different mammalian genomes are investigated with respect to their transcript diversity. The main focus of the thesis is investigation of how this transcript diversity reflects on the protein structures. Within the three genomes, specifically Transcription Factor genes are analyzed. The methodologies employed include biological data retrieval from contemporary biomedical resources, storage of data in a relational database and further computational analyses.

Our results revealed that both in human and in mouse more than half of the TF genes analyzed have unique transcripts which code for proteins with unique domains. That is they have at least 2 unique transcripts coding for differential protein domain structures. Importantly, the unique domain coded by one of the TF transcripts and not the other conveys DNA-binding ability. This is the case for 51% of TF human genes and 52% of TF mouse genes. Given the lesser number of transcripts sequenced per rat TF genes in general, this percentage stays at 37%, as expected.

The overall conclusion from this thesis is that the majority of TF genes have transcript diversity and that this transcript diversity brings diversity in protein structures and thus in functions.

Keywords:

Transcription factor, genomes, transcripts, protein structure, domain function, DNA-binding, biological databases, data retrieval and storage.

ÖZ

Bu çalısmada, üç farklı memeli organizmanın genomlarının transkript çeşitliliği incelenmiştir. Tezin temel amacı ise transkript çeşitliliğinin protein yapısındaki etkilerini incelemektir. İncelenen üç genomun özellikle Transkripsiyon Faktörlerini (TF) kodlayan kısımları analiz edilmiştir. Bu çalışma süresince kullanılan metotlardan bazıları güncel biyomedikal kaynakları kullanarak biyolojik veri toplamak, toplanan veriyi saklamak, veriye çeşitli yollardan ulaşılabilecek bir bilgisayar veritabanına kaydetmek ve çeşitli hesaplamalı analizler yapmak olmuştur.

Sonuçlar göstermiştir ki hem insan hem de fare genomlarında analiz edilen TF genlerinin en az yarısı kendilerine özgü yapısal bölümler (domain) içeren proteinleri kodlayan transkriptlere sahiptir. Diğer bir deyişle, bu genlerin, her birinin farklı özgün yapısal bölümlere sahip en az 2 proteini kodlayan değişik transkriptlere sahip olduğu anlaşılmıştır. En önemlisi, iki TF transkriptinden sadece birinde gözlenmiş olan özgün yapısal bölümün, DNA ile bağ kurma kabiliyetine sahip olmasıdır. Bu farklılık insan TF genlerinin %51'inde, fare TF genlerinin ise %52'sinde gözlemlenmiştir. Sıçan TF genlerinin sekanslanmış transkript sayısının düşük olduğunu göz önüne alırsak, bu yüzdelik beklendiği gibi %37 civarında kalmıştır.

Bu tezden çıkarabileceğimiz genel sonuç şudur ki TF genlerinin çoğunun transkript çeşitliliği yüksektir, ve bu çeşitilik proteinlerde görülebilir yapısal ve buna bağlı olarak fonksiyonel farklılıklara yol açar.

Anahtar Kelimeler

Transkripsiyon faktörü, genom, transkriptler, protein yapısı, domain fonksiyonu, DNA-bağlaması, biyolojik veri, veri toplanması ve depolaması.

To my Love and my Parents

ACKNOWLEDGMENT

First of all, I would like to thank my supervisor, Assoc. Prof. Dr. Ekrem Varoğlu, and my co-supervisor, Assoc. Prof. Dr. Bahar Taneri, for their kindness, supervision, understanding, help and guidance throughout this study. Their encouragement made me interested in bioinformatics.

Especially, I am deeply grateful to my beloved husband, Pejman, for him endless love and being there for me when I need him the most.

I would like to extend my appreciation to my parents who have supported me emotionally, financially and morally.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Background

Transcript diversity is important in generating protein diversity and increasing the complexity, hence functionality of genomes. In this study, the focus is on three mammalian genomes; human, mouse, rat, their transcript diversity and the effect of this diversity on their protein structures. In particular, the transcription factor (TF) genes within the three genomes are studied.

The transcripts coded by each TF gene which each genome are analyzed with respect to the protein domains they code. Differential protein domain coding by different transcripts of the same gene is documented as an indicator of protein functional diversity.

TFs are required for the regulation of gene expression and they are found in all eukaryotic species. The number of TFs found within an organism rises with genome size [1] [2]. For example, in the human genome approximately 2600 proteins have DNA-binding domains, and most of these proteins are presumed to function as transcription factors [3]. Hence, approximately 10% of genes in the genome code for TFs [4], which makes this family, the single largest family of human proteins with a very important cellular function. Furthermore, previous studies have shown the TF protein structure variation due to transcript variation [5].

1.2 Thesis Contribution

In this study, the association between transcript diversity and protein domains is investigated. The work done includes analysis of different human, mouse and rat RNA isoforms coded by the same gene, which potentially produce proteins with different domain architectures and hence functionality. Similar work has been performed before in mice, demonstrating such differences [5].

1.3 Thesis Outline

At the outset, an overview of bioinformatics is introduced, and some molecular biology concepts which are very useful to understanding thesis are presented, along with a literature review in Chapter 2. Chapter 3 shows the methodology used to retrieve data and to design the database, as well as codes that were developed to analyze the TF genes which produce multiple transcripts. In Chapter 4 conclusion on the results and future works related to this field are provided.

Chapter 2

OVERVIEW of BIOINFORMATICS and MOLECULAR BIOLOGY

2.1 An Overview of Bioinformatics

Bioinformatics is an interdisciplinary field that develops and applies computational technologies to study biomedical questions [6]. Bioinformatics tools are used to manage, search and analyze large amounts of data (also referred to as “big data”) in the life sciences. As a methodology, bioinformatics is a top-down, holistic, data-driven, genome-wide and systems-wide approach that generates new hypotheses, finds new patterns, and discovers new functional elements [6][7].

The interdisciplinary nature of bioinformatics is reflected in that it studies questions in biology and medicine, while developing and applying methods in computer sciences, mathematics, statistics, and physics. It has some overlaps with medical/clinical informatics, systems biology, and synthetic biology.

The “bio” in bioinformatics signifies the biological questions it studies, many of them could be grouped under the conceptual framework from genotype to phenotype. Figure 2.1 is showing bioinformatics research from genotype to phenotype.

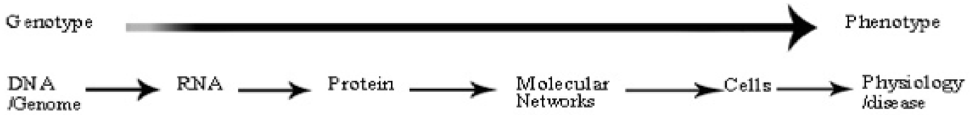


Figure 2.1: Bioinformatics research from the main theme of the Central Dogma and axis from genotype to phenotype. (The figure is taken from[8]).

The “informatics” in bioinformatics signifies the information processing and computational methods, and runs along the axis from data to discovery. Figure 2.2 is showing bioinformatics research from angle of information sciences.

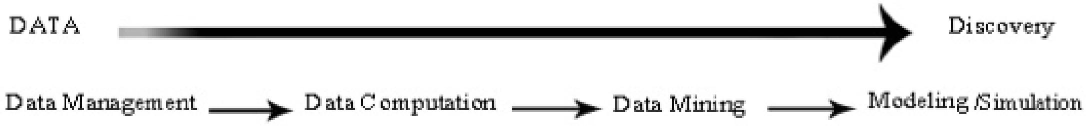


Figure 2.2 : Bioinformatics research from the angle of information sciences, from the main theme of from data to discovery. (The figure is taken from [9]).

During the last 60 years, bioinformatics has been rapidly developing, which is closely related to the developments of molecular biology and computer sciences. In 1950s and 1960s, many critical concepts and technologies in molecular biology were established. At the same time, many important concepts, software, and hardware of computer sciences were also generated. As it came to 1970s and 1980s, molecular biology and computer sciences started to merge, and this has been ongoing with increasing growth since 1990s [7][10].

Some of the classic bioinformatics questions first emerged around 1960s [7]. In the 1980s, the scientific questions, technologies, and research reached a critical mass, and bioinformatics as a field emerged, and experienced astonishing growth since the 1990s. The first appearance of the word “Bioinformatics” was in a little known Dutch paper published in 1970 [7][11]. In 1978 Pauline Hogeweb wrote in an English paper that she identified her research as in ”Bioinformatics” Many people refer to this paper as the origin of the word of ”Bioinformatics” [7][12].

2.2 Basic Molecular Biology Concepts

In this section some basic concepts of molecular biology are introduced.

2.2.1 DNA Structure

The structure of the DNA molecule was first inferred by James Watson and Francis Crick based primarily on X-array crystallography data collected by Maurice Wilkins and Rosalind Franklin, and chemical analysis of base composition of DNA conducted by Irwin Chargaff that known as Chargaff's rule [13-15]. According to this rule, adenine in one strand only hydrogen bonds with thymine, and guanine only hydrogen bonds with cytosine. [14]. The Chargaff's low is illustrated in Figure 2.3.

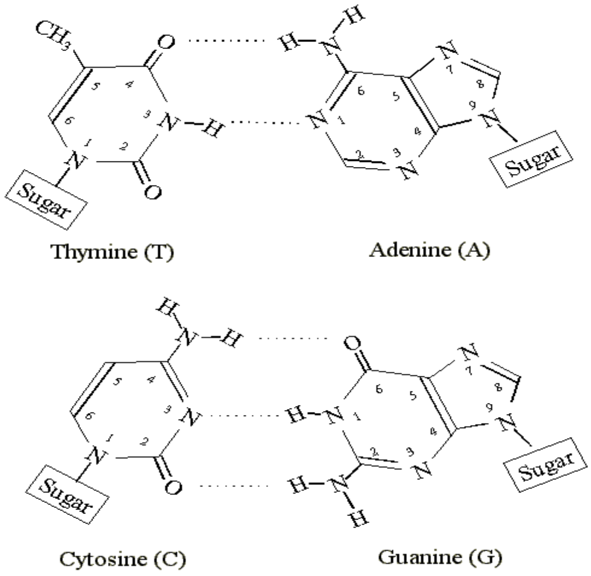


Figure 2.3: Chargaff's Law: A=T, G=C. (The figure is taken from [16]).

The key features of the structure are its right-handed double helical structure. Each helix consists of an alternating sugar-phosphate backbone with nitrogen bases projection toward the interior of each helix. One complete 360-degree turn of the helix covers 10 bases of length and equals 3.4 nanometers in physical distance along the axis of the molecule. The width of the double helix is 2 nanometers [13]. The DNA structure is shown in Figure 2.4.

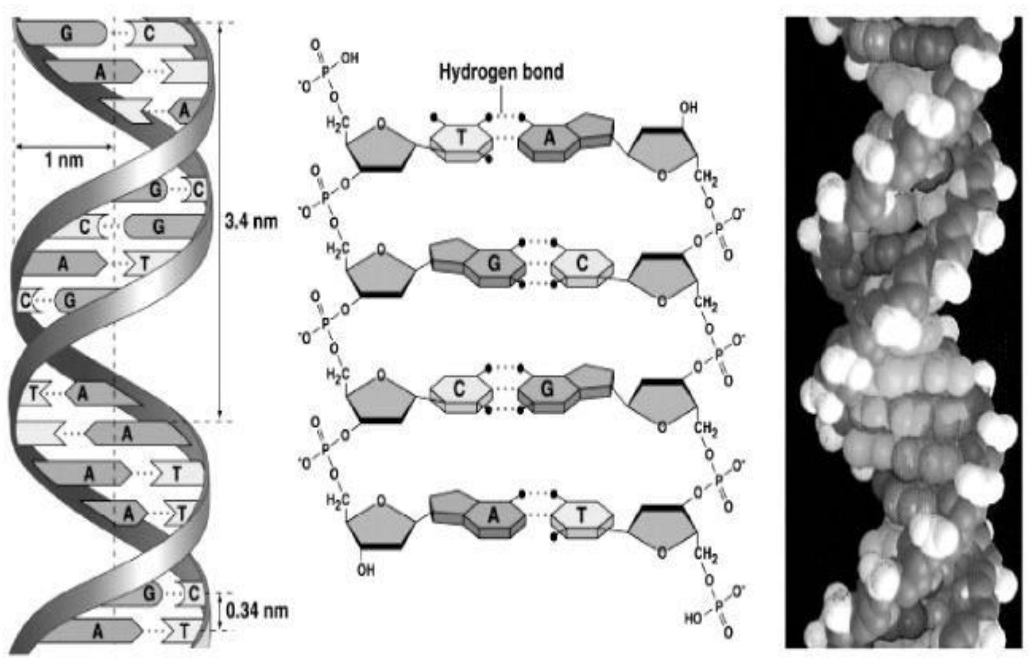


Figure 2.4: DNA structure. (The figure is taken from [17]).

The nucleotide bases are attached inside each backbone of the molecule so that the nucleotides in one helix or strand are hydrogen bonded to the bases in the other helix or strand. The hydrogen bonds hold the two strands of the double helix together. Guanine-cytosine base pairs form 3 hydrogen bonds while adenine-thymine base pairs form 2 hydrogen bonds. This makes guanine-cytosine base pairs more stable than adenine-thymine base pairs. Nucleotide pairing between strands also allows the sequence in one strand to determine the sequence in the complementary strand [18].

The two ends of a strand are not identical. One end of each strand a 3 prime hydroxyl group of the deoxyribose sugar is not involved in the backbone or it is free, while at the other end of the same strand the 5 prime hydroxyl group of the deoxyribose sugar at the end is free or may contain a phosphate that is free and not bonded to another deoxyribose sugar. This dissimilarity of the two ends of a strand creates the ability to uniquely distinguish each end of the strand. Because of this polarity of each strand the two strands of DNA are oriented in opposite directions or they are antiparallel [18].

2.2.2 Gene Expression

The central dogma of molecular biology describes two major steps: transcription and translation. These two steps are separated in eukaryotic cells [19]. Transcription occurs only within the nucleus to produce a pre-mRNA molecule. Eukaryotic mRNAs are modified before they are translated. Introns are removed and the remaining exons are spliced together. A 5' cap and a 3' tail are added. The processed mRNA travels to the cytoplasm where translation occurs [18][19]. These processes are shown in Figure 2.5.

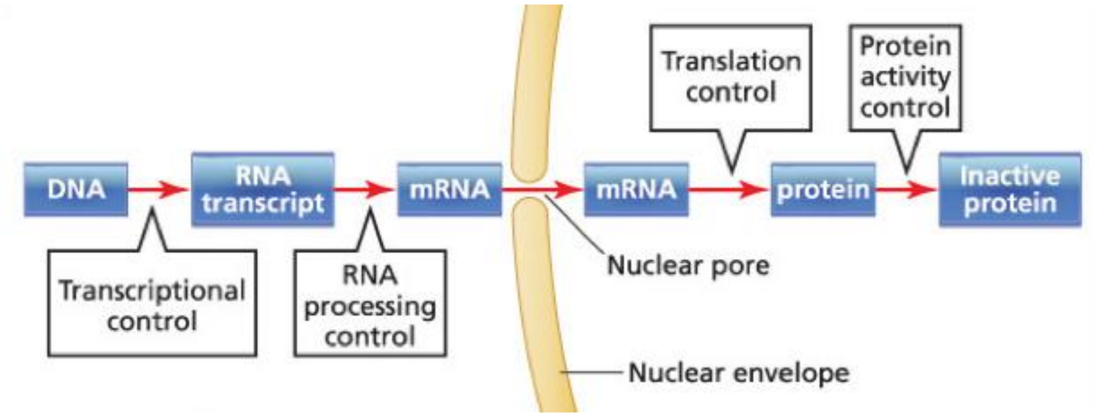


Figure 2.5: Control of Gene Expression in Eukaryotes. (The figure is taken from [20]).

The sequence of nucleotide bases in DNA carries genetic information in units that are referred to as genes. Structural genes encode the information for specific proteins. These genes are composed of numerous short-coding sequences referred to as exons, interspersed between long stretches of noncoding sequences referred to as introns [18].The structure of a eukaryotic gene is illustrated in Figure 2.6.

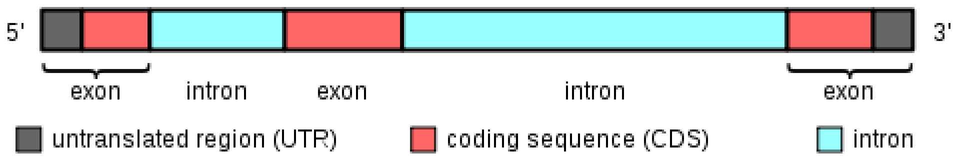


Figure 2.6: Basic structure of a eukaryotic gene. (The figure is taken from [21]).

To create a protein, a gene must first be transcribed into a sequence of nucleotide bases in form of a messenger RNA (mRNA) molecule [18].

Firstly, the genetic information in cells from DNA is read and transcribed into a pre-mRNA molecule. Mature mRNA is produced from pre-mRNA by RNA processing, this process includes capping, splicing, and polyadenylation of the transcript [22]. Then mRNA provides the code to construct a protein by a process referred to as translation. The mRNA sequence is then translated into an amino acid sequence of a protein [18].

This sequence of amino acids in a protein molecule determines the shape and chemical characteristics of the protein. Thus, each gene specifies a specific protein in the cell that carries out a specific function based on its chemical characteristics and molecular shape. This function of the specific protein gives the cell and the organism the specific trait coded for by the gene [23]. It is interesting that one gene could code for more than one type of mRNA molecule and hence could result in different protein products. This protein diversity generated by the transcript diversity is the focus of this thesis, as further described in the following section.

2.2.2.1 Transcription

Transcription is the synthesis of messenger RNA. The process of transcription has three stages: initiation, elongation, and termination [24].

A structural gene is constituted of a sequence of bases in a DNA molecule consisting of a coding region with an upstream promoter and a terminator downstream of the coding region. Attachment of RNA polymerase to the promoter region and formation of an open complex, starts transcription. But, for RNA polymerase to successfully attach to a eukaryotic promoter and make the transcription begin, a set of proteins referred to as transcription factors (TFs) should first assemble on the promoter [18][25]. Initially, proteins called basal factors bind to a short sequence in the promoter called the TATA box. Later on other basal proteins bind to form the full transcription factor complex, which is now able to recruit the RNA polymerase. Another set of transcription factors called co-activators link the basal factors with activators. Activators are regulatory proteins, they have the ability to bind DNA sequences called "enhancers". Many enhancers, which are scattered around the chromosome, could bind different activators, which provide a variety of responses to various signals. When a second kind of regulatory protein referred to as repressor binds to a "silencer" sequence located near to or overlapping an enhancer sequence, the corresponding activator can no longer bind DNA. After this process, RNA polymerase binds to promoter and initiates transcription [25-26].

RNA polymerase moves along the template strand of the DNA, synthesizing the complementary single-strand messenger RNA molecule. Synthesis is in the 5' to 3' direction, with new nucleotides being added to the 3' end of the growing messenger RNA molecule. As the RNA polymerase advances along the DNA, it unwinds a new stretch of DNA and allows the previous stretch to close [27]. The messenger RNA sequence is elongated as the RNA polymerase moves down the DNA molecule, until the RNA polymerase reaches the terminator region. When sequences in the terminator region are encountered, transcription is terminated. In fact, when RNA polymerase reaches a specific sequence of nucleotides on the DNA referred to as the transcription terminator, a hairpin loop structure forms in the messenger RNA causing the RNA polymerase and the messenger RNA to dissociate from the DNA. This causes RNA polymerase to dissociate from the DNA molecule, and the completed transcript is released [27-28]. The main stages of transcription mechanism are shown in Figure 2.7.

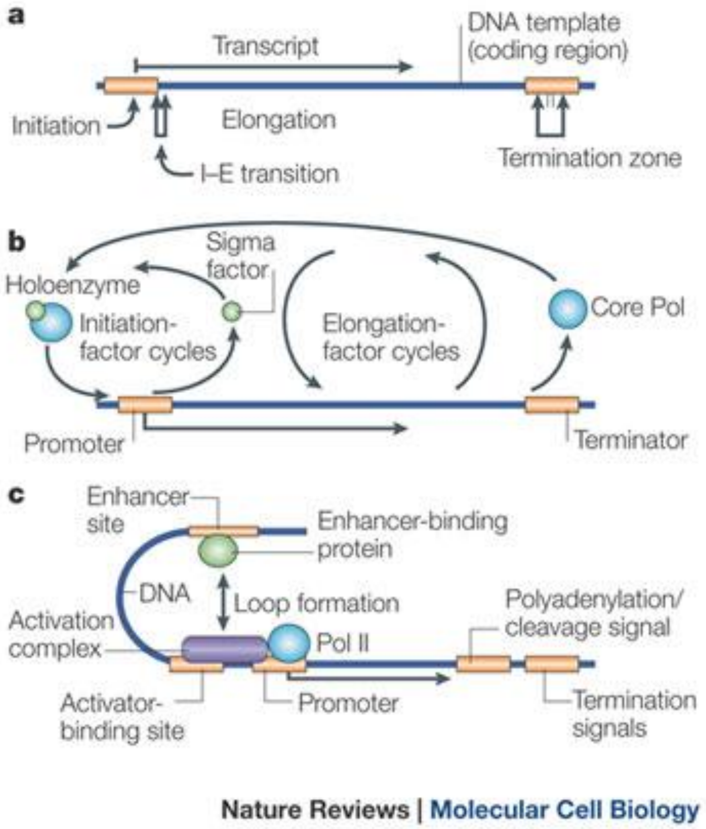


Figure 2.7: Phases of eukaryotic transcription. (The figure is taken from [29]).

2.2.2.2 Translation

Translation begins when messenger RNA binds to the ribosome. The initial transfer RNA (tRNA) occupies the P site on the ribosome [30]. Subsequent tRNAs with bound amino acids, first enter the ribosome at the A site, as sown in Figure 2.8.

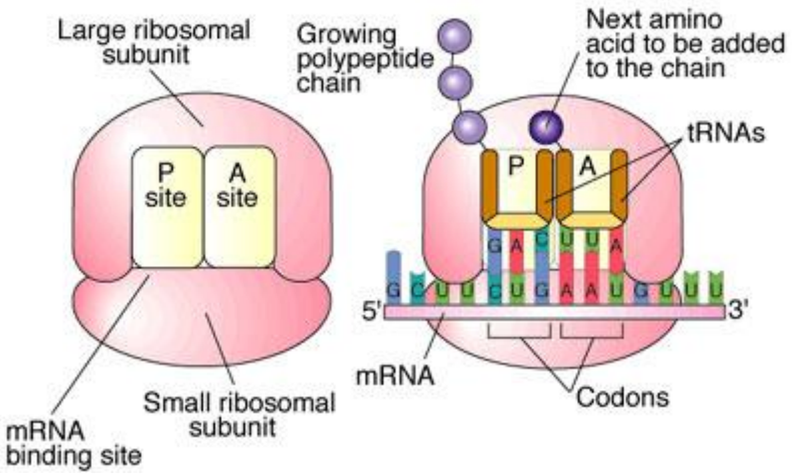


Figure 2.8: Ribosome structure. (The figure is taken from [31]).

The complementary matching of three nucleotides on the transfer RNA, called the anticodon, and three nucleotides on the messenger RNA, called the codon, ensures the correct sequence of amino acids. The messenger RNA passes along the ribosome in short spurts of 3 nucleotides at a time. As this occurs, the initial transfer RNA is moved to the E site and its amino acid is transferred to the second amino acid at the P site. At the same time, a new codon is presented at the A site. The initiating transfer RNA, which now no longer carries an amino acid, leaves the E site and the next transfer RNA, with a complementary anticodon, enters the A site. Each time a new codon sequence moves into the A site, a new transfer RNA brings in an amino acid. The old transfer RNA paired with the previous codon is passed to the P site and then to the E site as the amino acid it carried is transferred to the growing amino acid chain. As the ribosome proceeds down the messenger RNA a stop codon is finally encountered. At this point the ribosomal complex falls apart and the protein is released into the cell [30-31]. Translation proceeds in three phases. The first phase is, initiation, during which the ribosome is bound to the specific initiation (start) site on the mRNA. The second phase, elongation, consists of joining amino acids to the growing polypeptide chain according to the sequence specified by the message. The termination codon gives the signal for the third and last stage of protein synthesis, which is termination [32]. The main stages of translation mechanism are shown in Figure 2.9.

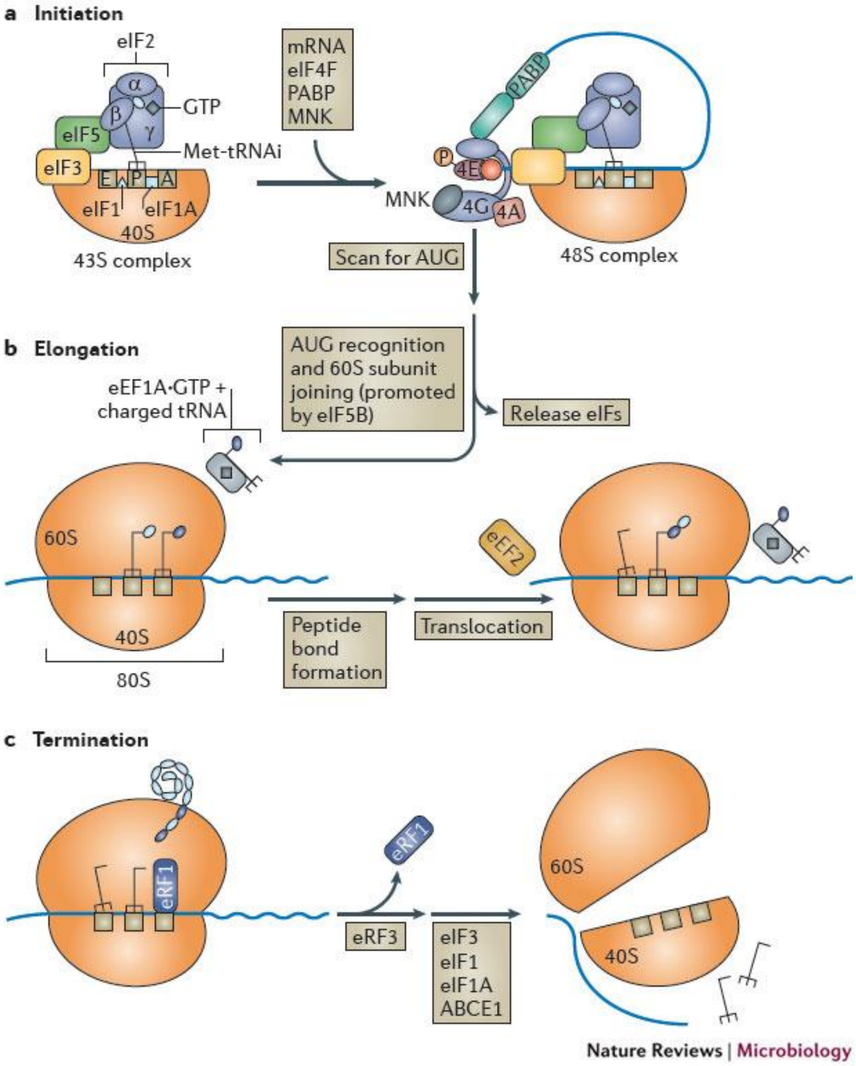


Figure 2.9: Stages of eukaryotic translation. (The figure is taken from [33]).

2.2.3 RNA Splicing

Most eukaryotic genes are consisted of numerous short-coding sequences referred to as exons, interspersed between long stretches of noncoding sequences referred to as introns [34]. RNA splicing removes introns from the pre-mRNA and attaches the exons together. Splicing involves a complex referred to as the spliceosome that has subunits referred to snRNPs. Each snRNP contains a small nuclear RNA and proteins. Specific sequences are essential for intron removal by the spliceosome.

Among the requirements are a GU at the 5' end of the intron (also referred to as the 5' splice site) and AG at the 3' end (or 3' splice site). A branch site toward the middle of the intron is also needed, this sequence contains an adenine (A) that plays an important role in the intron removal [35-36]. The 5', 3' splice sites and the branch site are shown in Figure 2.10.

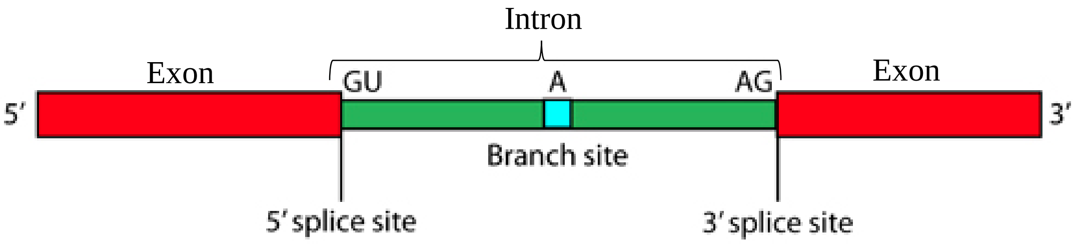


Figure 2.10: Sequences required for splicing. (The figure is taken from [37]).

Splicing involves several detailed steps. Firstly U1 snRNP binds to the 5' splice site and later on U2 snRNP binds to the branch site. Figure 2.11 shows these initial reactions [35-37].

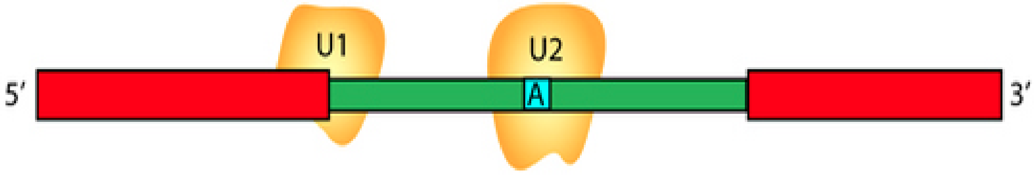


Figure 2.11: Binding of U1 and U2 snRNPs to the pre-mRNA molecule. (The figure is taken from [37]).

Next, the trimer of U4, U5 and U6 snRNPs binds, completing the spliceosome assembly [35-37]. The spliseosome assembly is shown in Figure 2.12.

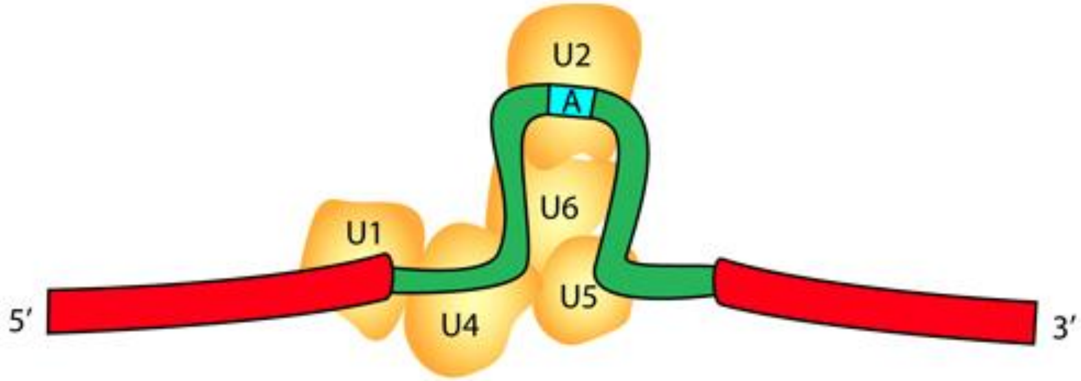


Figure 2.12: Spliceosome assembly. (The figure is taken from [37]).

The 5' splice site is cut, and the 5' end of the intron is attached to the adenine in the branch site to form a structure referred to as the lariat. Then the U1 and U4 snRNPs are released, and the U6 and U5 snRNPs shift positions and finally the 3' splice site is cut and the exons are connected together, meanwhile the lariat is released along with the parts of the spliceosome which remained [35-36]. The spliceosome disassembly is shown in Figure 2.13.

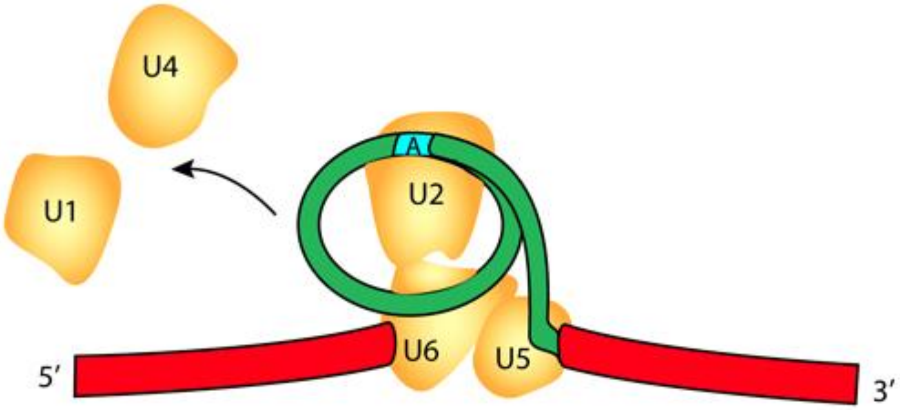


Figure 2.13: Spliceosome disassembly. (The figure is taken from [37]).

The spliceosome subunits will later dissociate from the lariat, and the lariat will be degraded. The final outcome is that two exons have been covalently attached to each other, and the intervening intron has been removed [35]. Figure 2.14 shows the exons connected together.

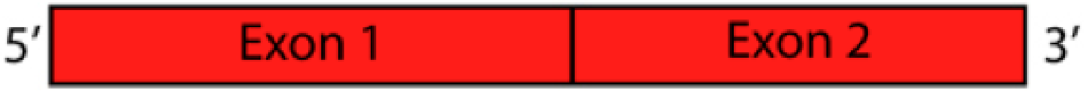


Figure 2.14: The exons connected to each other. (The figure is taken from [37]).

2.2.4 Alternative Splicing

The process that the primary transcript of a gene is reorganized in different ways to produce different transcripts is called alternative splicing [38]. By differential use of exons and introns, various transcripts with different nucleotide sequences could be generated with the alternative splicing mechanism. As a result, the sequence of the amino acids produced from the same gene but different transcripts could result in different protein sequences, and hence potentially different protein structures [38]. Alternative splicing has been observed as a mechanism to produce tissue, specific proteins from a single gene. Depending on the tissue, different proteins could be produced in different tissues from a single gene. This process could be thought of a multiplication process that increases the possible proteins that are produced from a single gene and overall from one genome [39].

Alternative splicing is a major source of protein diversity in living organisms. It has been estimated that at least 70% of all genes in the human genome are alternatively spliced and this number expands continuously [40]. The alternative splicing mechanism is exemplified in Figure 2.15.

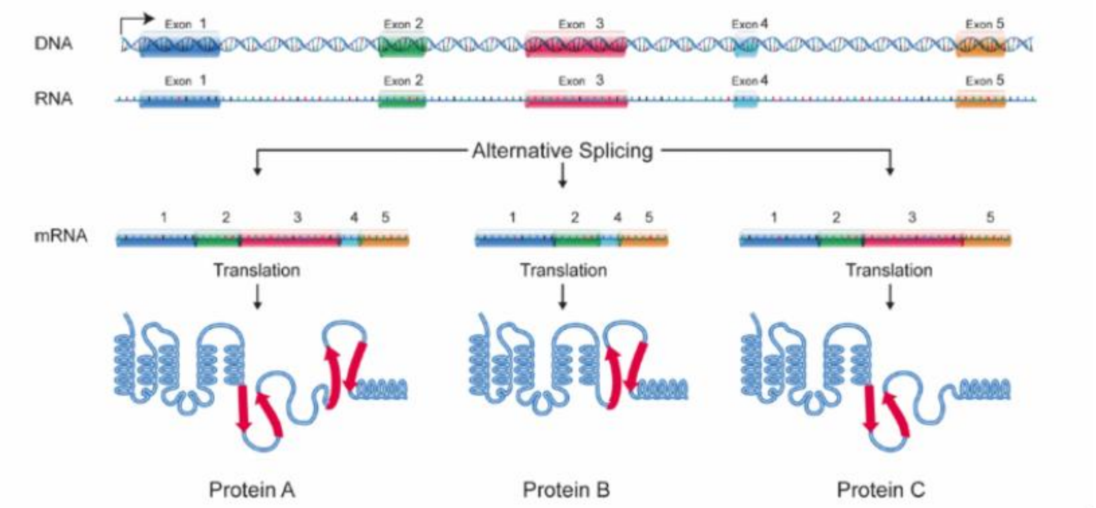


Figure 2.15: Example of alternative splicing mechanism. (The figure is taken from [41]).

2.2.4.1 Types of Alternative Splicing

The different types of alternative splicing [42] are as follows (Figure 2.16):

a. Alternative promoter selection: A different promoter is used for different splice variants. This results in a different start of the mRNA transcript.

b. Alternative selection of cleavage/polyadenylation sites: Different exons are spliced based on recognition of different cleavage or polyadenylation sites, entire exons could be skipped. This results in a different exon at the 3ʹ end of the transcript.

c. Intron retention: Introns are used as coding regions. A sequence that is normally considered an as intron is retained in the final transcript that serves as a template for translation.

d. Cassette exons: Entire exons could be skipped in the middle of the protein, resulting in a different transcript.

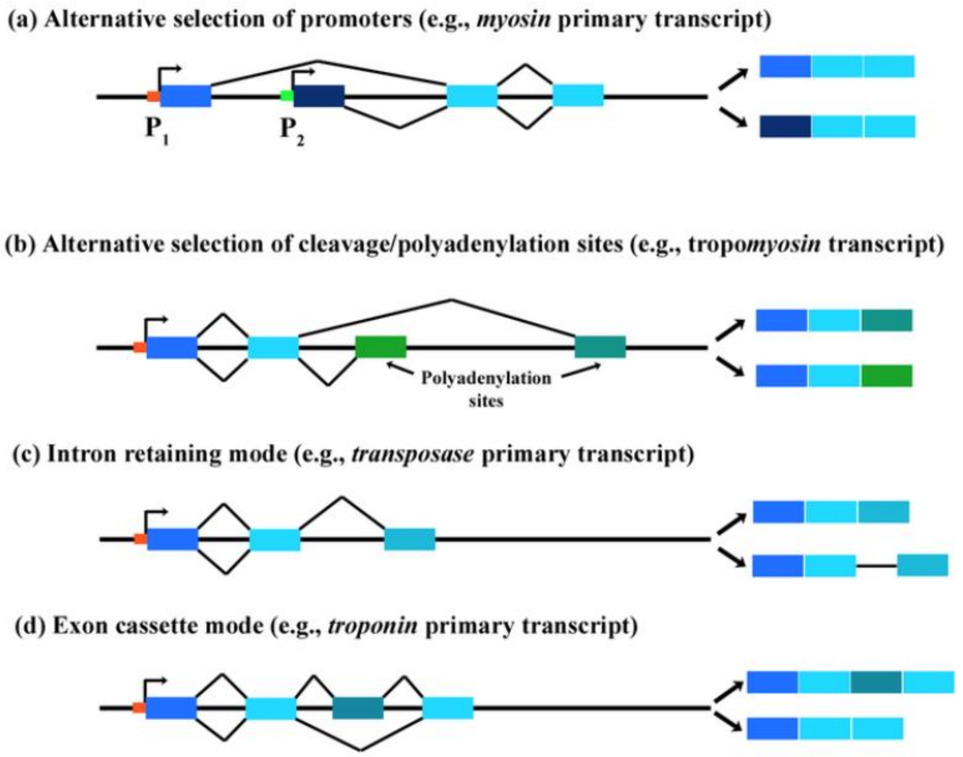

(a) Alternative selection of promoters (e.g., *myosin* primary transcript)
P1
P2
(b) Alternative selection of cleavage/polyadenylation sites (e.g., tropo*myosin* transcript)
Polyadenylation sites
(c) Intron retaining mode (e.g., *transposase* primary transcript)
(d) Exon cassette mode (e.g., *troponin* primary transcript)

Figure 2.16: Different types of alternative splicing. (The figure is taken from [43]).

2.2.5 Protein

Proteins are polymers. A polymer is any molecule that is made up individual building blocks that are linked together. The individual building blocks are called monomers. The monomers that make up proteins are called amino acids. A chain of amino acids is called a polypeptide. Polypeptide is a chain of three or more amino acids that are linked together, which is not yet folded. Protein is a polypeptide that has folded into a 3-dimentional shape. Ultimately, proteins are made of two or more polypeptides [18]. Figure 2.17 shows an amino acids sequence forming a short polypeptide chain.

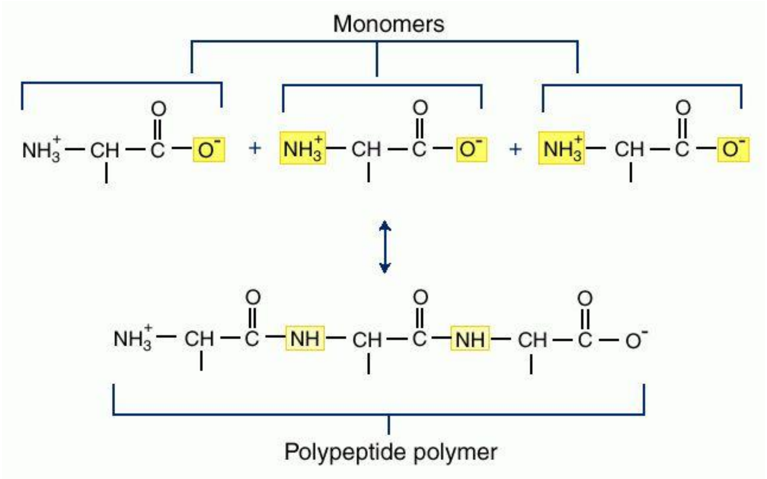


Figure 2.17: A short amino acid sequence. (The figure is taken from [44]).

2.2.5.1 Structure of Amino Acids

The structure of a typical amino acid consists of an amino group (NH2). At the other hand, there is a carboxyl group (COOH). In addition, there is a central carbon atom, also known as the alpha (α) carbon which links together the amino group with the carboxyl group [18]. A hydrogen atom is bonded with this central carbon atom. Central carbon also binds a side chain, another atom or a group of atoms known as the R group (or side chain or variable group). The general structure of an amino acid is shown in Figure 2.18.

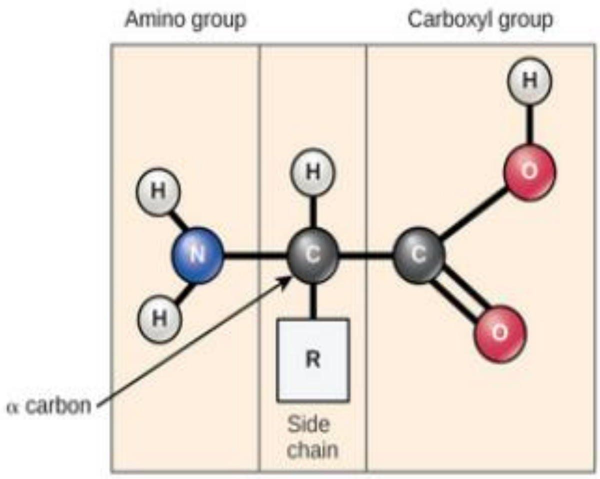
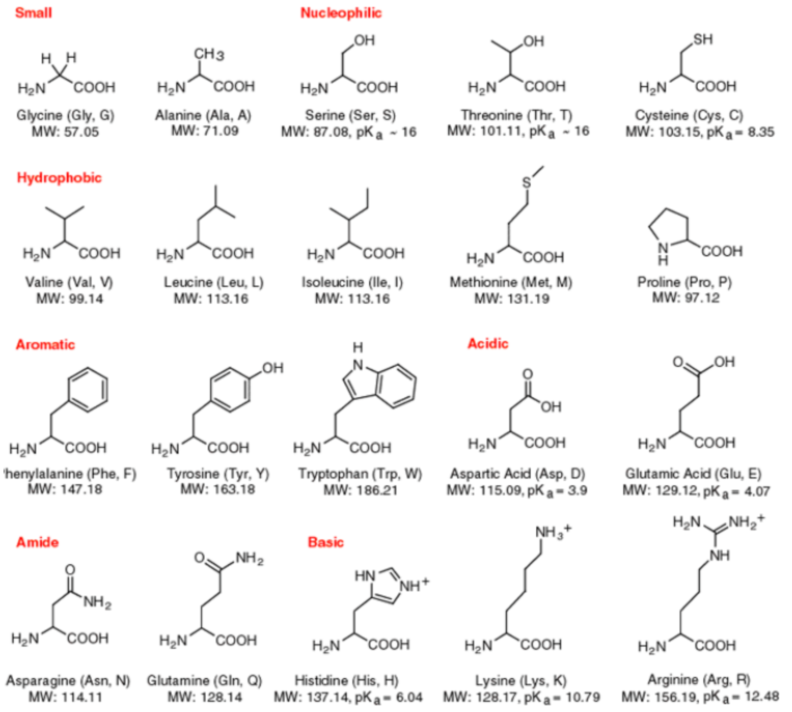


Figure 2.18: The structure of an amino acid. (The figure is taken from [45]).

For each amino acid, the R group (or side chain) is different. Different amino acids have different variable groups. The chemical nature of the side chain identifies the nature of the amino acid its function and properties [18]. Figure 2.19 shows all amino acid types.



Small
Nucleophilic
Glycine (Gly, G) MW: 57.05
Alanine (Ala, A) MW: 71.09
Serine (Ser, S) MW: 87.08, pKa ~ 16
Threonine (Thr, T) MW: 101.11, pKa ~ 16
Cysteine (Cys, C) MW: 103.15, pKa = 8.35
Hydrophobic
Valine (Val, V) MW: 99.14
Leucine (Leu, L) MW: 113.16
Isoleucine (Ile, I) MW: 113.16
Methionine (Met, M) MW: 131.19
Proline (Pro, P) MW: 97.12
Aromatic
Acidic
henylalanine (Phe, F) MW: 147.18
Tyrosine (Tyr, Y) MW: 163.18
Tryptophan (Trp, W) MW: 186.21
Aspartic Acid (Asp, D) MW: 115.09, pKa = 3.9
Glutamic Acid (Glu, E) MW: 129.12, pKa = 4.07
Amide
Basic
Asparagine (Asn, N) MW: 114.11
Glutamine (Gln, Q) MW: 128.14
Histidine (His, H) MW: 137.14, pKa = 6.04
Lysine (Lys, K) MW: 128.17, pKa = 10.79
Arginine (Arg, R) MW: 156.19, pKa = 12.48

Figure 2.19: The 20 types of amino acid. (The figure is taken from [46]).

Multiple amino acids can be linked together to create a polypeptide through a reaction known as condensation reaction or dehydration reaction. A condensation reaction removes a molecule of water(H_2O) in the making of a bond. Then, the carbon of carboxyl group and the nitrogen of amino group are linked together to create a peptide bond. A peptide bond is a simple type of covalent bond that links together two amino acids [18]. Figure 2.20 shows the peptide bond.

The protein's shape, size, and function depends on the sequence and the number of its amino acids [24].

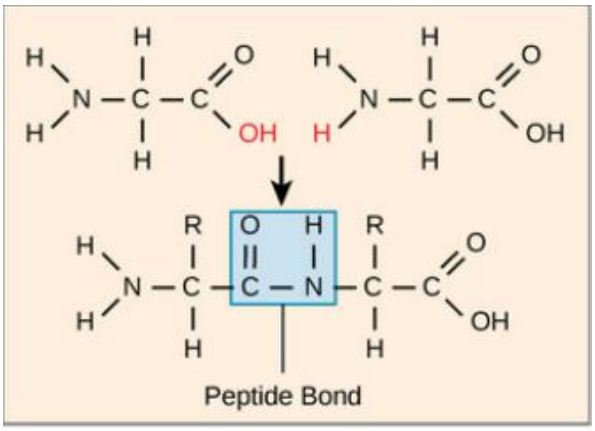


Figure 2.20: Peptide bond. (The figure is taken from [45]).

The products formed by such linkages are also referred to as peptides. For understanding how a protein reaches its final form or final structure, four levels of the protein structure: primary, secondary, tertiary, and quaternary should be analyzed [18].

2.2.5.2 Primary Structure

The primary structure simply is the order of amino acids that make up the polypeptide chain [47]. It is the sequence of how these amino acids are linked together. The primary structure is held together with the peptide bond this is a type of covalent bond that links amino acids together [18]. Figure 2.21 shows the primary protein structure.

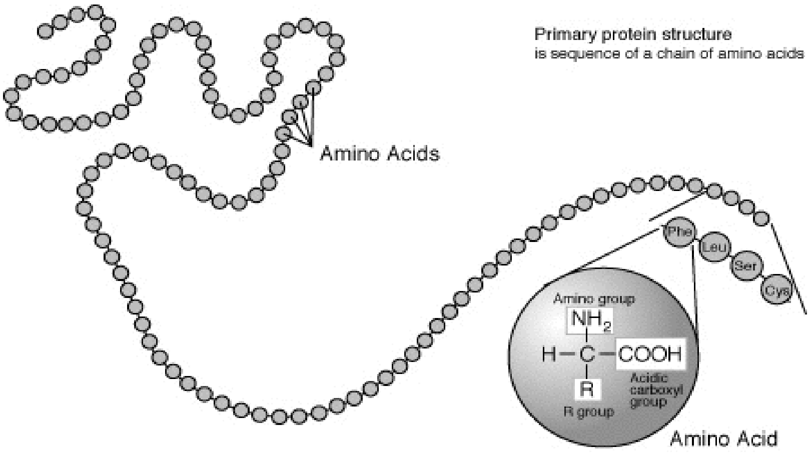


Figure 2.21: Primary protein structure. (The figure is taken from [48]).

2.2.5.3 Secondary Structure

The secondary protein structure is the hydrogen-bonding pattern of the peptide backbone of the protein [49]. The most common secondary structures are α-helix and β-pleated sheet [18][45][50]. The backbone is formed as a helix. The α-helix is one segment of the chain that starts forming helical structure [50][52]. The β-pleated sheet is the chain of amino acids that may consist of parallel strands, antiparallel strands or a mixture of parallel and antiparallel strands. The secondary structures, α-helix and β-pleated sheets are held together through hydrogen bonding [18]. Figure 2.22 shows the secondary protein structure.

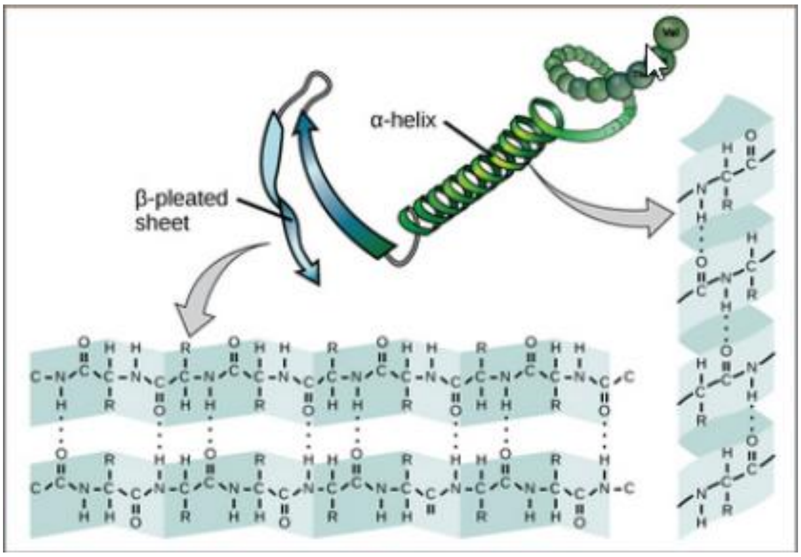


Figure 2.22: Secondary protein structure. (The figure is taken from [51]).

2.2.5.4 Tertiary Structure

The tertiary structure is a three-dimensional structure of entire polypeptide chain, which forms partly become of the chemical interactions of the polypeptide chain. In particular, interactions between the R groups generate the tertiary structure. The tertiary structure is held together through many interactions [45]. Firstly, hydrogen bonds between the different variable groups of amino acids are among these interactions. Some amino acids can interact through ionic bonds, Van der Waals interactions and lastly via disulfide bridges. These are different type of bonds that could be found in the tertiary protein structures [18][45][52]. Figure 2.23 shows the tertiary protein structure.

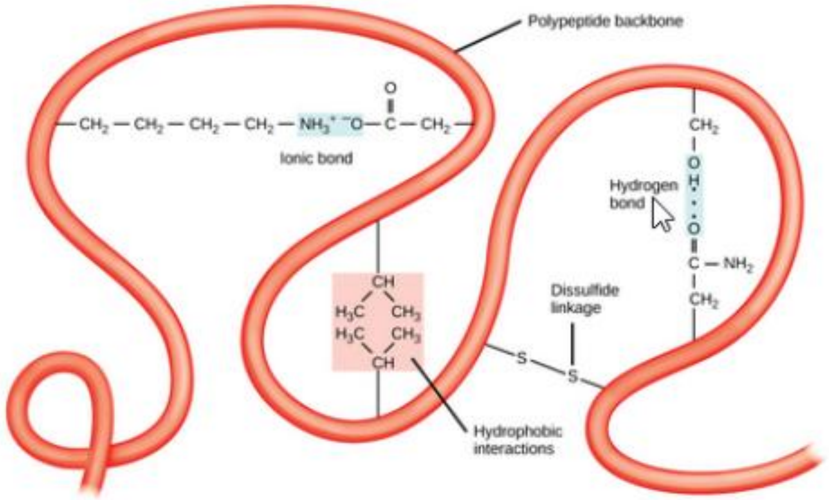


Figure 2.23: Tertiary protein structure. (The figure is taken from [45]).

2.2.5.5 Quaternary Structure

Not all proteins have quaternary structure, when there are more than one polypeptide chain making up a particular protein, a quaternary structure could form [45]. They interact together and form a fully functional protein. The four levels of protein structure are illustrated in Figure 2.24.

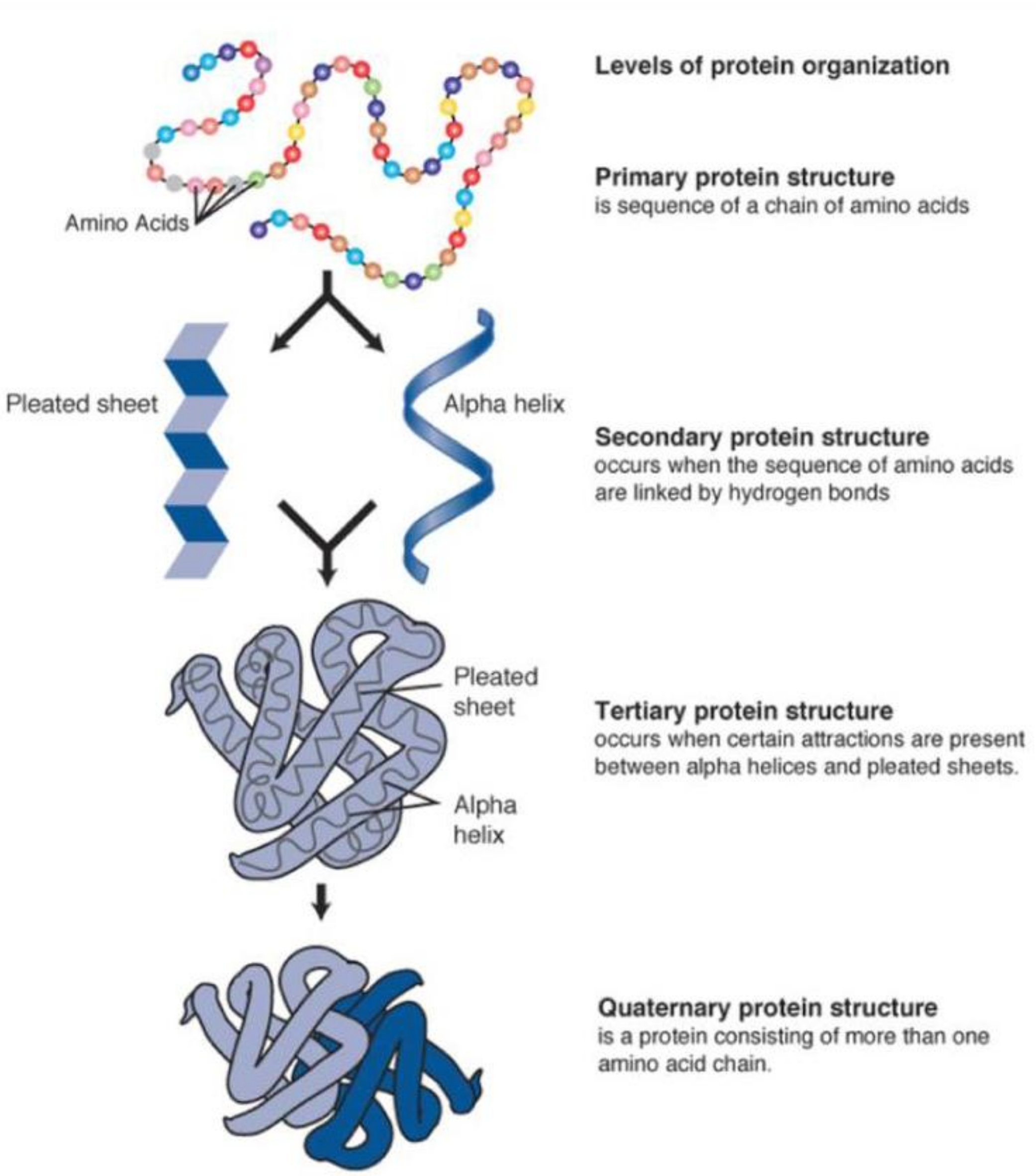

Levels of protein organization
Amino Acids
Primary protein structure
is sequence of a chain of amino acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids
are linked by hydrogen bonds
Pleated
sheet
Alpha
helix
Tertiary protein structure
occurs when certain attractions are present
between alpha helices and pleated sheets.
Quaternary protein structure
is a protein consisting of more than one
amino acid chain.

Figure 2.24: The four levels of protein structure. (The figure is taken from [45]).

2.2.5.6 Protein Domain

Domains are parts of a protein with specific functions and structures. Protein domains encode portions of proteins and can be assembled together to form translational units, a genetic part spanning from translational initiation to translational termination [53].

Proteins are divided into different categories according to sequence or structural similarity. Proteins can be divided into different categories based on [53]:

- the FAMILIES they belong to.
- the DOMAINS they contain.
- the SEQUENCE FEATURES they possess.

Domains could be termed as units within a protein with specific structural characteristics and functions. In general, a domain is responsible for a distinct function of a protein or an interaction. Put together, different domains of a protein generate its overall function. One domain could be found in different proteins with variety of functions [54].

2.2.6 Source of RNA Transcript Diversity

RNA transcript diversity evolves from several different mechanisms, including RNA splicing, this mechanism removes introns from the pre-mRNA and attaches the exons together. As discussed previously Alternative splicing is a major source of transcript diversity in living organisms. Alternative transcription initiation and polyadenylation site usage, RNA editing and trans-splicing over long distances from different gene loci are among the other mechanisms generate transcript diversity [55-56].

2.2.7 Source of Protein Diversity

Three main molecular mechanisms are considered to contribute expanding the repertoire and diversity of proteins present in living organisms: first, at DNA level (gene polymorphisms and single nucleotide polymorphisms); second, at messenger RNA (pre-mRNA and mRNA) level including alternative splicing (also termed differential splicing or cis-splicing). Finally, at the protein level protein diversity is mainly driven through Post-translational Modification (PTM) and specific proteolytic cleavages [56-57].

Chapter 3

METHODOLOGY

The goal of this thesis is to investigate the association between transcript diversity and protein diversity coded by TF genes in 3 different genomes. In order to achieve this goal data must first be collected from relevant biological databases. The data retrieved is stored in a relational database in order to avoid redundancy and allow easy analysis. Finally, statistical analysis of the data stored is performed in order to obtain the results.

3.1 Biological Databases and Resources Used

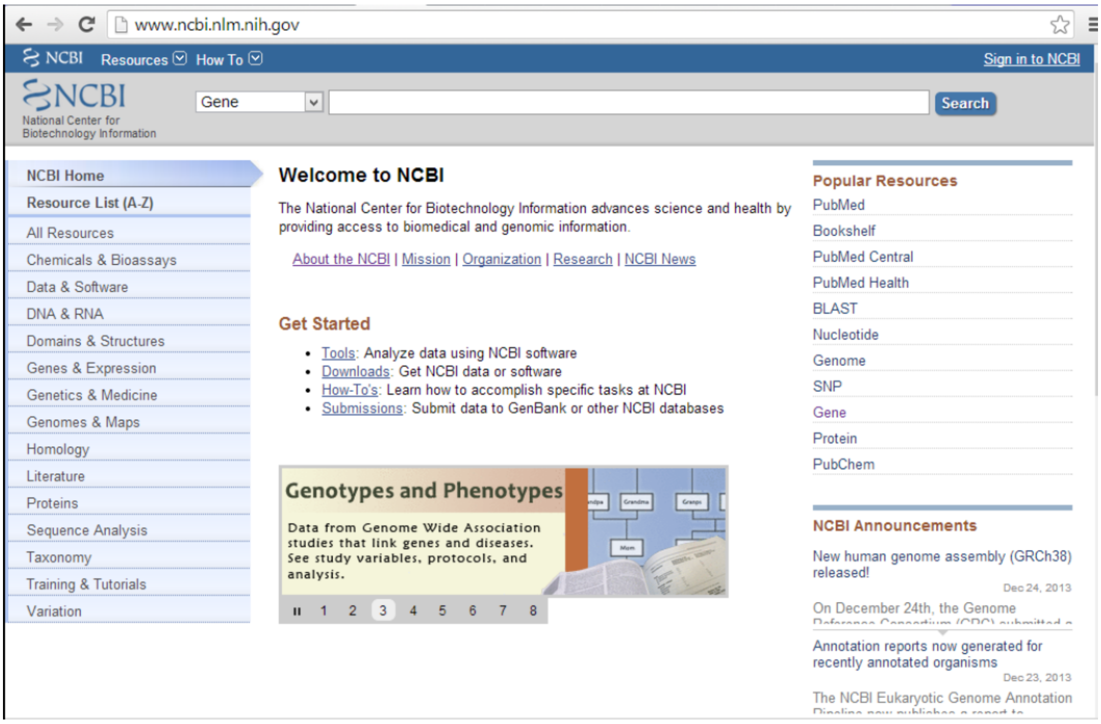
Several of the most frequently used biological databases and resources are NCBI [58], Ensembl [59], BioMart [60], SMART [61] and Pfam [62]. These resources contain several different levels of information for DNA, RNA, protein domains and structures. In the following sections, further detailed information about these databases is presented.

3.1.1 National Center for Biotechnology Information (NCBI)

One of the largest centralized bioinformatics resources is maintained by the National Center for Biotechnology Information (NCBI) at the National Institute of Health (NIH) in the US. NCBI contains many database resources including information for DNA, RNA, and proteins (domains and structures), expression data, variations, literature and etc. In addition, software tools for data retrieval and analysis are provided. All the databases are available online through the Entrez search engine [63].

As of end of 2013, over 1000 complete whole genome sequences are available from the NCBI Genome resource. NCBI also has a resource called Gene which integrates various useful information about each genome. As of end of 2013, NCBI Gene resource provides annotations for about 14 million genes in 11,000 species [63].

The Entrez global query is an integrated search and retrieval system that provides access to all databases simultaneously with a single query string and user interface. Entrez can efficiently retrieve related sequences, structures, and references [64]. The screenshot of the NCBI web homepage is provided in Figure 3.1.



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | NCBI News

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Data from Genome Wide Association studies that link genes and diseases. See study variables, protocols, and analysis.

NCBI Announcements

New human genome assembly (GRCh38) released!
Dec 24, 2013

On December 24th, the Genome Reference Consortium (GRC) submitted a

Annotation reports now generated for recently annotated organisms
Dec 23, 2013

The NCBI Eukaryotic Genome Annotation Pipeline now publishes a report to

Figure 3.1 : Homepage of NCBI. (The figure is taken from [58]).

3.1.2 Ensembl

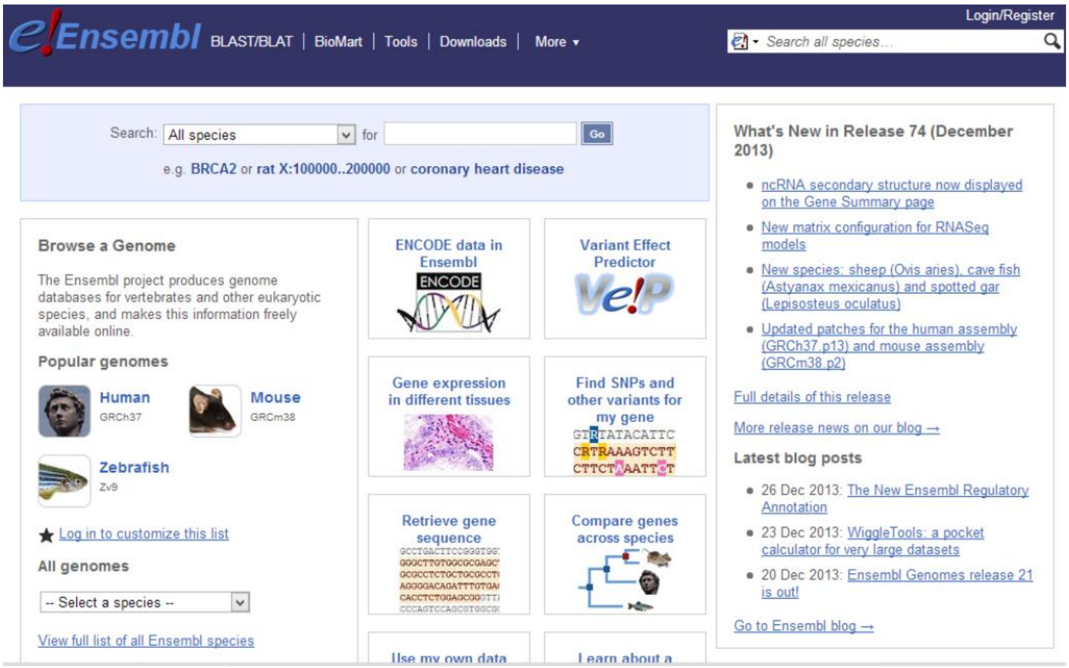
An important resource at the European Bioinformatics Institute (EBI) is the Ensembl database which is a comprehensive database for gene and genome annotations. Ensembl provides comprehensive genome databases that incorporate many types of data and annotations in addition to the genomic sequences, including gene expression data, genetic variations, cross-species comparision, etc. It includes data for many vertebrates and other eukaryotic species. Over one hundred databases containing biological data are included in Ensembl [65].

The naming convention used for genes in Ensembl is shown in Table 3.1. The Ensembl identifiers are stable, which means that in a future update they refer to the same gene ids.

Table 3.1: Gene naming convention used for human species in Ensembl.

ENS**G**###	Ensembl Gene ID
ENS**T**###	Ensembl Transcript ID
ENS**P**###	Ensembl Protein ID
ENS**E**###	Ensembl Exon ID

For non-human species a suffix is added; for example, ENS**MUS**G###, is used for mouse. Information such as gene sequence, splice variants, and further annotation can be retrieved at the genome, gene and protein level. Ensembl genome browser is updated every two months [66]. Figure 3.2 shows the homepage of Ensemble.



e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

★ Log in to customize this list

All genomes

View full list of all Ensembl species

What's New in Release 74 (December 2013)

- ncRNA secondary structure now displayed on the Gene Summary page
- New matrix configuration for RNASeq models
- New species: sheep (Ovis aries), cave fish (Astyanax mexicanus) and spotted gar (Lepisosteus oculatus)
- Updated patches for the human assembly (GRCh37.p13) and mouse assembly (GRCm38.p2)

Full details of this release

More release news on our blog →

Latest blog posts

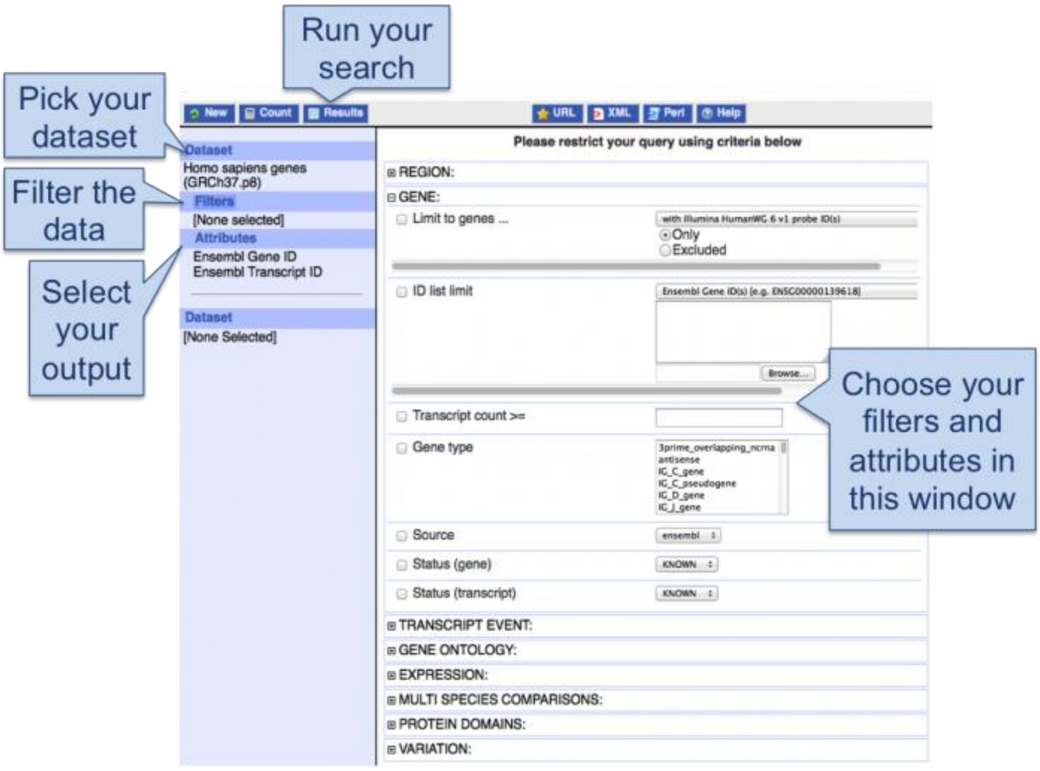
- 26 Dec 2013: The New Ensembl Regulatory Annotation
- 23 Dec 2013: WiggleTools: a pocket calculator for very large datasets
- 20 Dec 2013: Ensembl Genomes release 21 is out!

Go to Ensembl blog →

Figure 3.2: Home page of Ensembl genome browser. (The figure is taken from [59]).

3.1.3 BioMart

BioMart is a web interface used for retrieving data from Ensembl. Ensembl BioMart provides a comprehensive visualization for data access and querying. Ensembl BioMart is created by using the database schemas and data generated by the various components of the Ensembl project. It is comprised of seven databases, including, Ensembl Genes, Ensembl Variation, Ensembl Regulation. The Ensembl Genes database release 61 contains 52 fully supported species and the Ensembl Variation database contains data for 18 species [67]. A sample interface for BioMart is shown in Figure 3.3.



Dataset
Homo sapiens genes (GRCh37.p8)
Filters
[None selected]
Attributes
Ensembl Gene ID
Ensembl Transcript ID

Dataset
[None Selected]

Please restrict your query using criteria below

⊞ REGION:

⊟ GENE:

□ Limit to genes ... with Illumina HumanWG 6 v1 probe ID(s)
● Only
○ Excluded

□ ID list limit Ensembl Gene ID(s) [e.g. ENSG00000139618]

□ Transcript count >=

□ Gene type 3prime_overlapping_ncrna
antisense
IG_C_gene
IG_C_pseudogene
IG_D_gene
IG_J_gene

□ Source ensembl

□ Status (gene) KNOWN

□ Status (transcript) KNOWN

⊞ TRANSCRIPT EVENT:

⊞ GENE ONTOLOGY:

⊞ EXPRESSION:

⊞ MULTI SPECIES COMPARISONS:

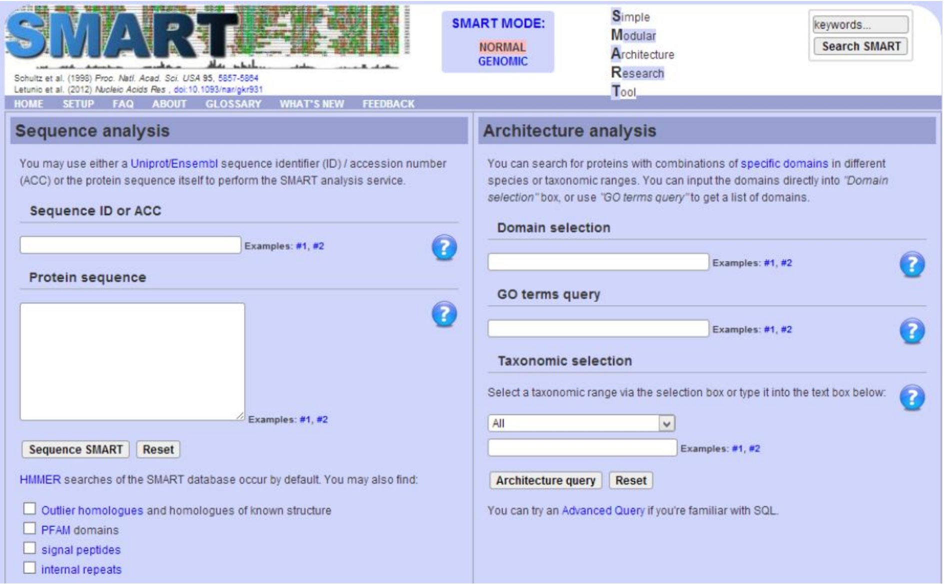
⊞ PROTEIN DOMAINS:

⊞ VARIATION:

Figure 3.3: A sample BioMart interface. (The figure is taken from [60]).

3.1.4 Simple Modular Architecture Research Tool (SMART)

SMART is a biological database, which is used for the identification and annotation of protein domains [68]. In order to analyze the domain architectures, SMART uses Profile-Hidden Markov Models (PHMM). It provides a platform for the comparative study of complex domain architectures in genes and proteins. The database is hosted by the European Molecular Biology Laboratory (EMBL) in Heidelberg. A protein domain in the SMART database has an ID consisting of the letters SM followed by a number. Some protein domains also have names. Figure 3.4 shows the homepage of SMART webpage.



Sequence analysis

You may use either a Uniprot/Ensembl sequence identifier (ID) / accession number (ACC) or the protein sequence itself to perform the SMART analysis service.

HMMER searches of the SMART database occur by default. You may also find:

□ Outlier homologues and homologues of known structure
□ PFAM domains
□ signal peptides
□ internal repeats

Architecture analysis

You can search for proteins with combinations of specific domains in different species or taxonomic ranges. You can input the domains directly into *"Domain selection"* box, or use *"GO terms query"* to get a list of domains.

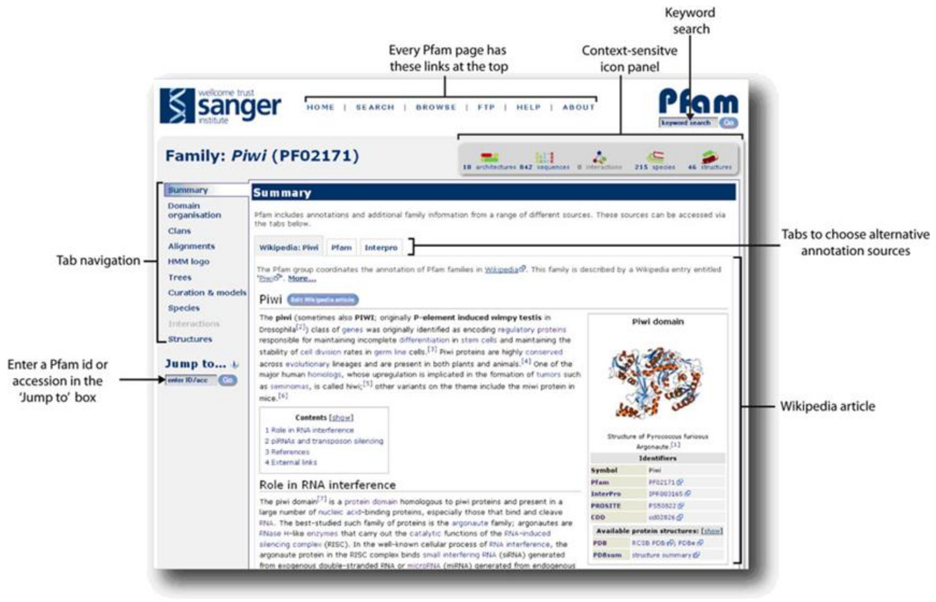
Select a taxonomic range via the selection box or type it into the text box below:

You can try an Advanced Query if you're familiar with SQL.

Figure 3.4 : Homepage of SMART webpage. (The figure is taken from [61]).

3.1.5 Protein Family (Pfam) Database

Pfam is a high-quality comprehensive database of multiple sequence alignments. It stores over 13,000 protein families, and many common protein domains. A protein family in the Pfam database has an ID consisting of the letters PF followed by a number. Some families also have names [69]. A typical Pfam family webpage is shown in Figure 3.5.



Keyword search

Every Pfam page has these links at the top

Context-sensitve icon panel

Tabs to choose alternative annotation sources

Tab navigation

Enter a Pfam id or accession in the 'Jump to' box

Wikipedia article

Family: *Piwi* (PF02171)

Summary

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

The Pfam group coordinates the annotation of Pfam families in Wikipedia. This family is described by a Wikipedia entry entitled "Piwi". More...

Piwi

The **piwi** (sometimes also **PIWI**; originally **P-element induced wimpy testis** in Drosophila[2]) class of genes was originally identified as encoding regulatory proteins responsible for maintaining incomplete differentiation in stem cells and maintaining the stability of cell division rates in germ line cells.[3] Piwi proteins are highly conserved across evolutionary lineages and are present in both plants and animals.[4] One of the major human homologs, whose upregulation is implicated in the formation of tumors such as seminomas, is called hiwi;[5] other variants on the theme include the miwi protein in mice.[6]

Contents [show]
1 Role in RNA interference
2 piRNAs and transposon silencing
3 References
4 External links

Role in RNA interference

The piwi domain[7] is a protein domain homologous to piwi proteins and present in a large number of nucleic acid-binding proteins, especially those that bind and cleave RNA. The best-studied such family of proteins is the argonaute family; argonautes are RNase H-like enzymes that carry out the catalytic functions of the RNA-induced silencing complex (RISC). In the well-known cellular process of RNA interference, the argonaute protein in the RISC complex binds small interfering RNA (siRNA) generated from exogenous double-stranded RNA or microRNA (miRNA) generated from endogenous

Structure of Pyrococcus furiosus Argonaute.[1]

Identifiers

Available protein structures: [show]

Figure 3.5: Typical Pfam family webpage. (The figure is taken from [62]).

3.2 Data Retrieval and Organization

The related data to TF genes in the human, mouse and rat species are extracted and stored in a relational database for further analysis. In the following sections, detailed information about these processes is presented. Figure 3.6 shows the E-utility tool is used to extract data from NCBI.

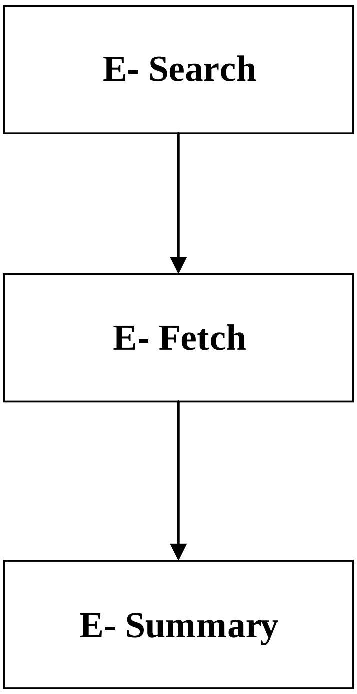


Figure 3.6: Flow diagram for Data retrieval using E-utilities.

3.2.1 Data Retrieval

Integrating data from multiple sources enhances research in bioinformatics. However, access to different resources and working with different file formats, which use various naming conventions, are not easy. One solution to this problem would be to provide links to other databases. For example, when a user searches for a particular gene, it should be possible to find the gene that encodes the protein sequence, protein families and protein domains.

In this thesis, mouse, rat and human TF genes and related information are examined. In order to collect data related to these species, firstly, data is retrieved from the NCBI database.

As an example the following query is used to search for the human entries related to "Transcription Factor" in Entrez Gene. The following query is used:

'((transcription factor) AND "Homo sapiens"[porgn]) AND "current only"[Filter]'

Information about the TF genes from other eukaryotes (mouse and rat) can be similarly obtained by modifying the above query as follow:

'((transcription factor) AND "Mus musculus"[porgn:__txid10090]) AND "current only"[Filter]'; for mouse, and

'((transcription factor) AND "Rattus norvegicus"[porgn:__txid10116]) AND "current only"[Filter]'; for rat.

Secondly, in order to retrieve data about each TF gene the E-utility tool of NCBI is used.

NCBI E-utilities, is the API to the Entrez system of databases. The E-utilities give code access to all of the major functions of Entrez, including text searching in databases, such as PubMed, Nucleotide, or Gene. Downloading records in various formats and linking between records in different databases are also possible. There are seven E-utility CGIs, all sharing the same base URL.

A Perl program is used to access data through the Entrez system. In particular, E-search utility is used to query each species. An example of a search query in perl is shown in the following.

```
$db=’gene’;
$query= ‘((transcription factor) AND “Rattus norvegicus”[porgn:_txid10116])
            AND “current only”[Filter]’;
```

E-fetch uses a query such as the one above to retrive from NCBI about each TF gene.

An example of E-search utility using a perl query is shown in the following.

```
$base=’http://eutils.ncbi.nlm.nih.gov/entrez/eutils/’;
$url=$base.”esearch.fcgi?db=$db&term=$query&usehistory=y”;
```

The UIDs retrived and stored on the history server and used to fetch records for each TF gene using E-fetch. An example of E-fetch query is shown in the following.

```
$base=’http://eutils.ncbi.nlm.nih.gov/entrez/eutils/’;
$url=$base.”efetch.fcgi?db=$db&id=$query&retmode=xml”;
```

The data retrieved by E-fetch contains multiple parts. The Ensembl Ids for each gene are contained in the “document summary” part. In order to access the summaries, the E-summary utility should be used. Since the data format is XML, it is required to have a XML interpreter. In order to deal with large list of UIDs two parameters are used in E-summary; A query key and a web environment string (WebEnv). Since the value of &usehistory is set to “yes”, in the E-search query, the returned E-summary will contain these two values. These parameters are used for retrieving the summary. An example of E-summary query is shown in the following.

```
$url=$base.”esummary.fcgi?db=$db&query_key=$key&WebEnv=$web”;
$docsum=get($url);
```

Perl programming language, uses XML::Twig in order to access the data which is retrieved in XML format. The XML tree structure is used to access the required parts of the data.

Finally, the aforementioned steps are used in the order presented above in order to retrieve Ensembl Transcription Factor Gene Ids for the three species. The number of Ensembl TF genes retrieved from NCBI database is shown in Table 3.2.

Table 3.2: Number of TF genes with respect to their species.

Species	Number of TF genes
Human	2152
Mouse	1567
Rat	1150

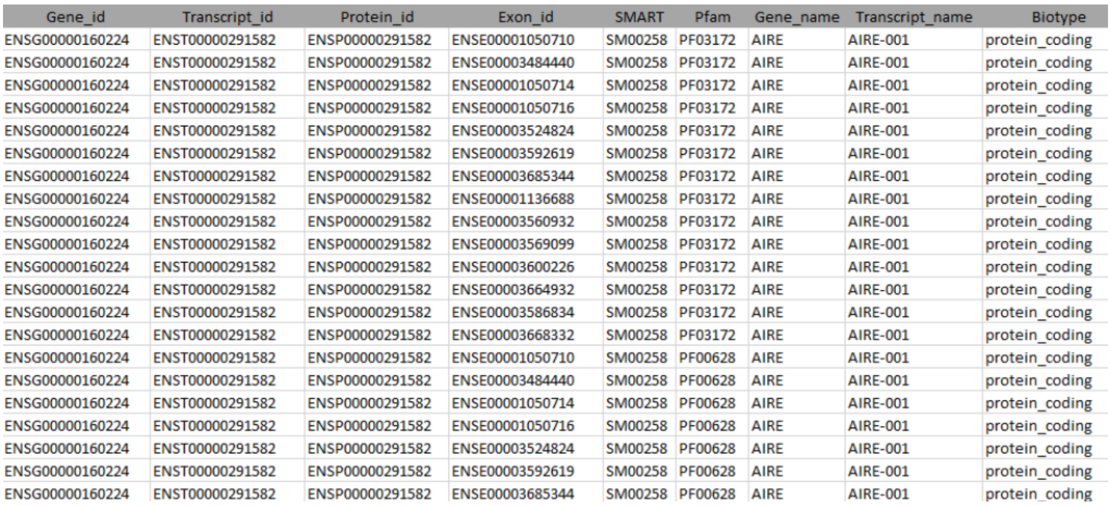
The data used in this thesis was collected in September 2013.

In the subsequent steps other attributes of TF genes which were obtained previously are extracted using BioMart. In order to run a BioMart query, firstly, a dataset is chosen among the three different species, *Homo sapiens*, *Mus musculus* and *Rattus* *norvegicus*. Secondly, filtering of data is done for a specific set of genes such as TF genes.

Finally, attributes such as Transcript Id, Protein Id, Exon Id, Association Gene Name, Biotype, Association Transcript Name, Description, SMART, and Pfam Domains are retrieved in order to determine the output columns.

3.2.2 Constructed Database

Efficient and proper storage of digital data retrieved is very important for further analysis. Earlier, the main way to store data on a computer was to store it in the form of files. However, file-processing systems have lots of disadvantages such as data redundancy and inconsistency, difficulty in accessing data, data isolation, difficulty in satisfying consistency, difficulty in ensuring database consistency, concurrent access by multiple users and security problems.



Gene_id	Transcript_id	Protein_id	Exon_id	SMART	Pfam	Gene_name	Transcript_name	Biotype
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00001050710	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003484440	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00001050714	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00001050716	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003524824	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003592619	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003685344	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00001136688	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003560932	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003569099	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003600226	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003664932	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003586834	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003668332	SM00258	PF03172	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00001050710	SM00258	PF00628	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003484440	SM00258	PF00628	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00001050714	SM00258	PF00628	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00001050716	SM00258	PF00628	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003524824	SM00258	PF00628	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003592619	SM00258	PF00628	AIRE	AIRE-001	protein_coding
ENSG00000160224	ENST00000291582	ENSP00000291582	ENSE00003685344	SM00258	PF00628	AIRE	AIRE-001	protein_coding

Figure 3.7: The output data of BioMart.

Database management systems (DBMS) are now used to solve most of the above problems. Underlying the structure of a database is the data model, which is a collection of conceptual tools for describing data, data relationships, data semantics, and consistency constraints [70]. The relational model is the most widely used data model. In the relational model the database is composed of a set of named relations or tables. Each relation contains a set of named attributes or columns and rows, which contain the value for each attribute. Each attribute has a domain [71-72].

The Entity-Relationship model (ER model) is a data model for describing a database. It is expressed in terms of entities, which are objects or concepts in the real world with an independent existence and can be differentiated from other objects. The relationships of entities are also represented [73-74]. The ER model is usually expressed in the form of an ER diagram.

Database normalization on the other hand is the process of organizing the tables of a relational database in order to minimize data redundancy and dependency [71][75]. In this thesis, the data retrieved from various biological databases is stored in a relational database. This data will later be used for further analysis.

The data is organized in different entities with respect to their semantic properties as shown in Figure 3.8.

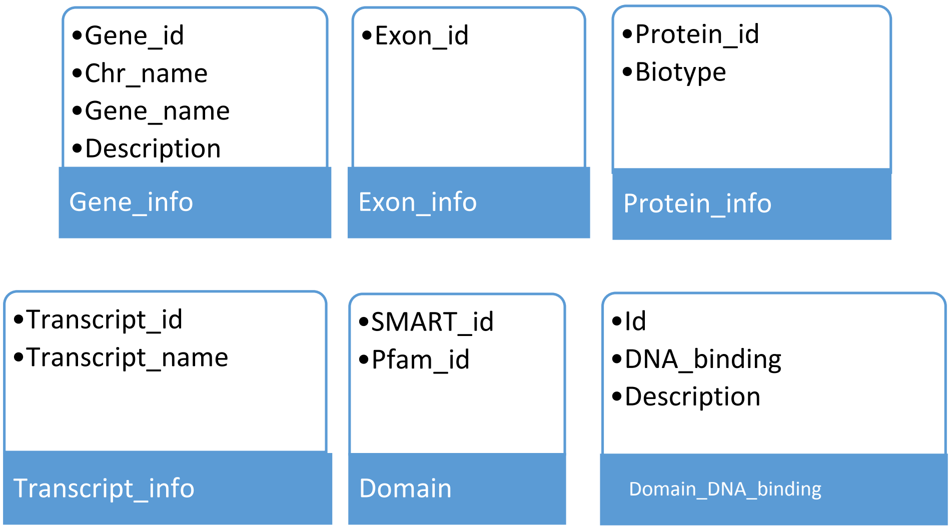


Figure 3.8: Entities used for storing the data.

The “Gene_info” entity contains “Gene_id”, “Gene_name”, “Chr_name” and “Description” attributes. The “Gene_id” is chosen as the primary key. The primary key should be unique and identify a specific record. The “Gene_id” attribute is used in Ensembl as an identifier for each TF gene, thus making it unique. The “Gene_name” attribute illustrates the name of each TF gene. This name is used in searching for TF genes with respect to their names. The “Chr_name” attribute shows the chromosome name in which this specified gene is located. The “Description” attribute specifies the function of this gene, its source and symbols.

The “Exon_info” entity contains only “Exon_id” attribute, which is also used as the primary key. The “Exon_id” shows the identifier of each exon which is used on the transcript sequence. Normally, other attributes regarding exons can be stored in this entity but such attributes are not used in this study. Hence, the table has only one attribute.

The “Protein_info” entity contains “Protein_id” and “Biotype” attributes. The “Protein_id” is chosen as the primary key. The “Protein_id” attribute specifies a unique sequence of amino acids, which are introduced as a protein. The “Biotype” attribute shows the gene type.

The “Transcript_info” entity contains “Transcript_id” and “Transcript_name”. The “Transcript_id” attribute is chosen as the primary key. The “Transcript_id” determines the specific transcript sequence. “Transcript_name” shows the name of each transcript.

The “Domain” entity contains “SMART” and “Pfam” attributes. Both of these attributes are defined as a composite primary key. The “SMART” attribute shows the identifier of domain in the SMART database and the “Pfam” attribute shows the identifier of the domain in the Pfam database.

The “Domain_DNA_binding” entity contains “Id”, “DNA_binding” and “Description” attributes. The “Id” field is defined as the primary key. This attribute shows the SMART or Pfam domain identifier. The “DNA_binding” attribute shows the function of this domain. The “Description” attribute contains a brief history of this domain. This entity is added after designing the database. It contains unique domains of TF genes, which produce two transcripts and their functions and descriptions. Figure 3.9 shows the E-R diagram for the designed database.

In the process of organizing the tables in a relational database the relationships between tables are defined. Large tables are divided into smaller tables or similar tables are joined.

Each TF gene contains multiple exons. Therefore, the relationship between these two entities is one to many (1:M). Each TF gene can produce one or more transcripts and

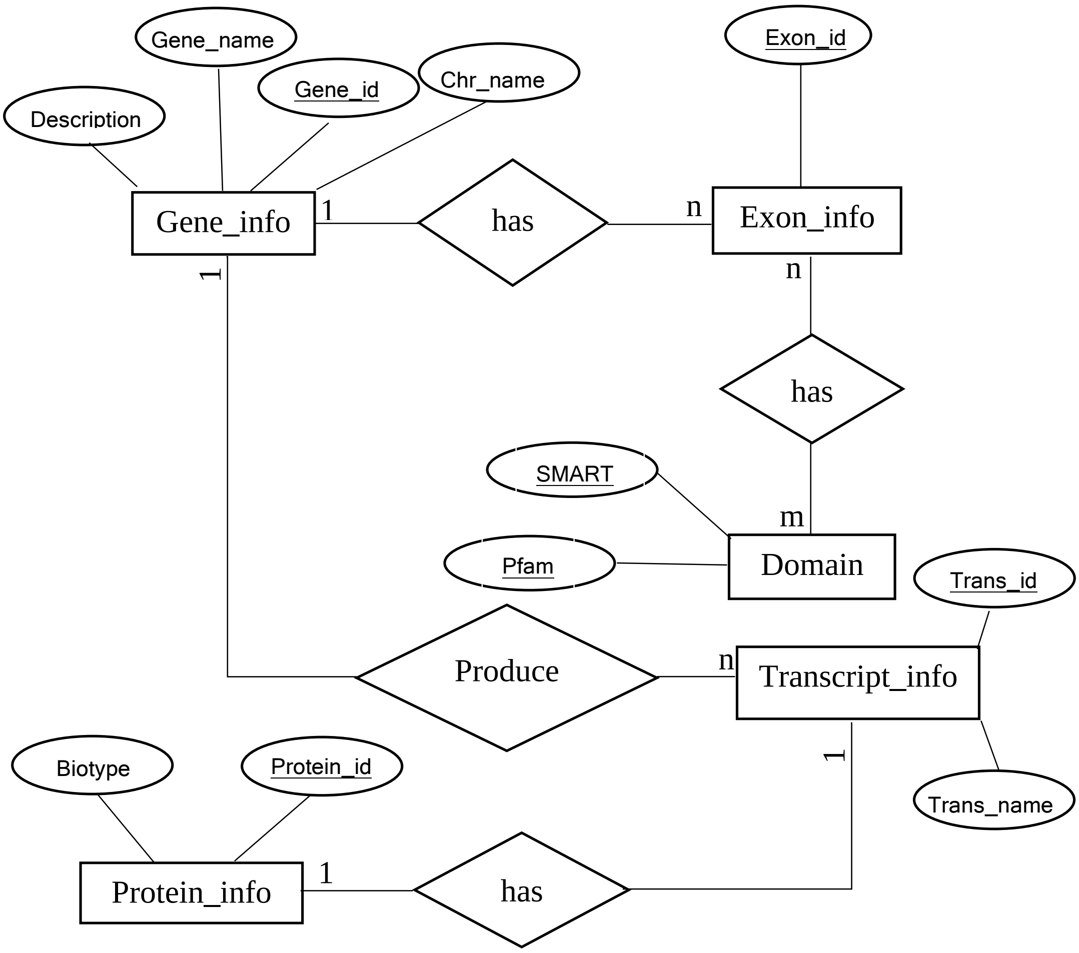


Figure 3.9: E-R diagram for the designed database.

each transcript produces one protein. So, the relationship between “Gene_info” and “Transcript_info” entities is one to many (1:M) and the relationship between “Transcript_info” and “protein_info” entities is one to one (1:1). The latter pair of entities form the relationship “Pro_trans_info” which as a relational table. The “Gene_info.Gene_id” attribute is a foreign key to the “Pro_trans_info” table.

Each exon may code for different domains and each domain may be coded by a different exon. Therefore, the “Exon_info” and “Domain” entities are related. The relationship between these two entities is many to many (N:M). A relational table is created which contains the primary key attributes of both “exon_info” and “Domain” entities. The name of this table is “Exon_domain_info”.

The primary keys of both “Exon_info” and “Domain” entities are present in this relational table. Since, both “Exon_info” and “Domain” tables contain only primary keys as attributes, there is no need to create an additional table. The “gene_id” from “Gene_info” and “Transcript_id” from “Pro_Trans_info” are specified as foreign keys to this table. The final database designed is shown in Figure 3.10.

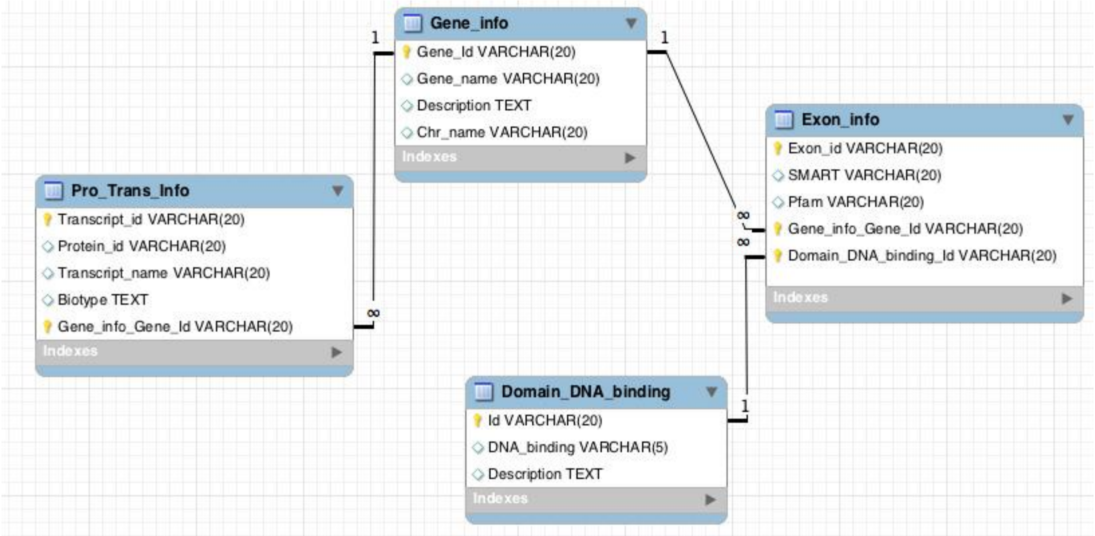
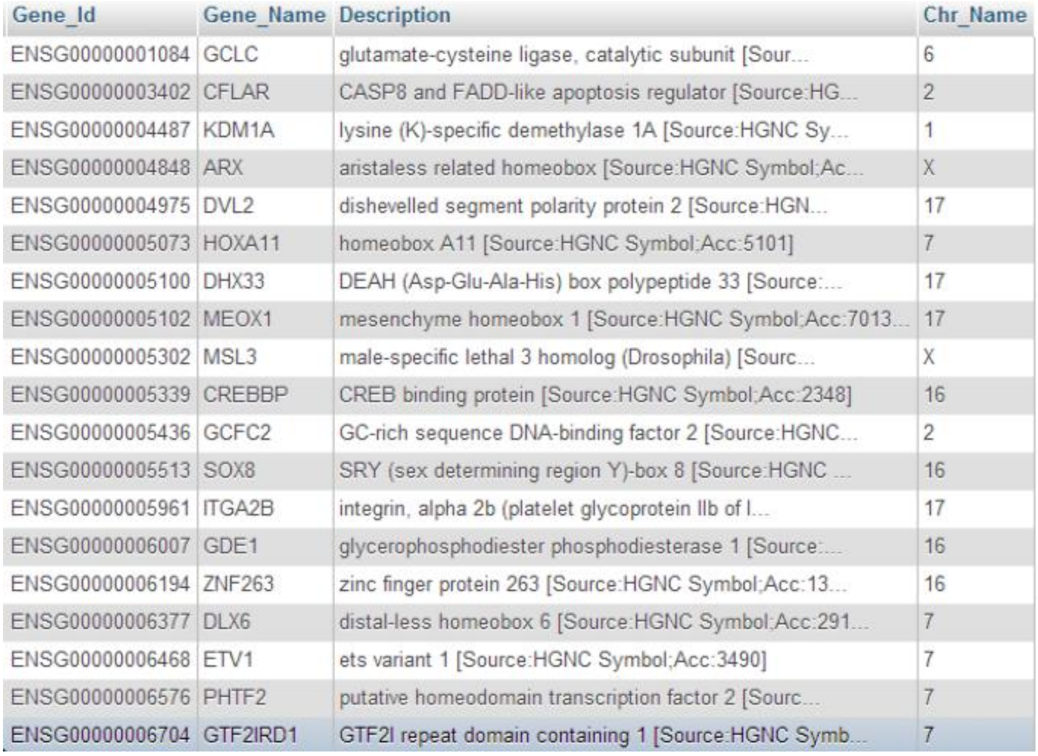


Figure 3.10: The relational database.

In this study, php my admin version 4.0.4 with mysql database is used.

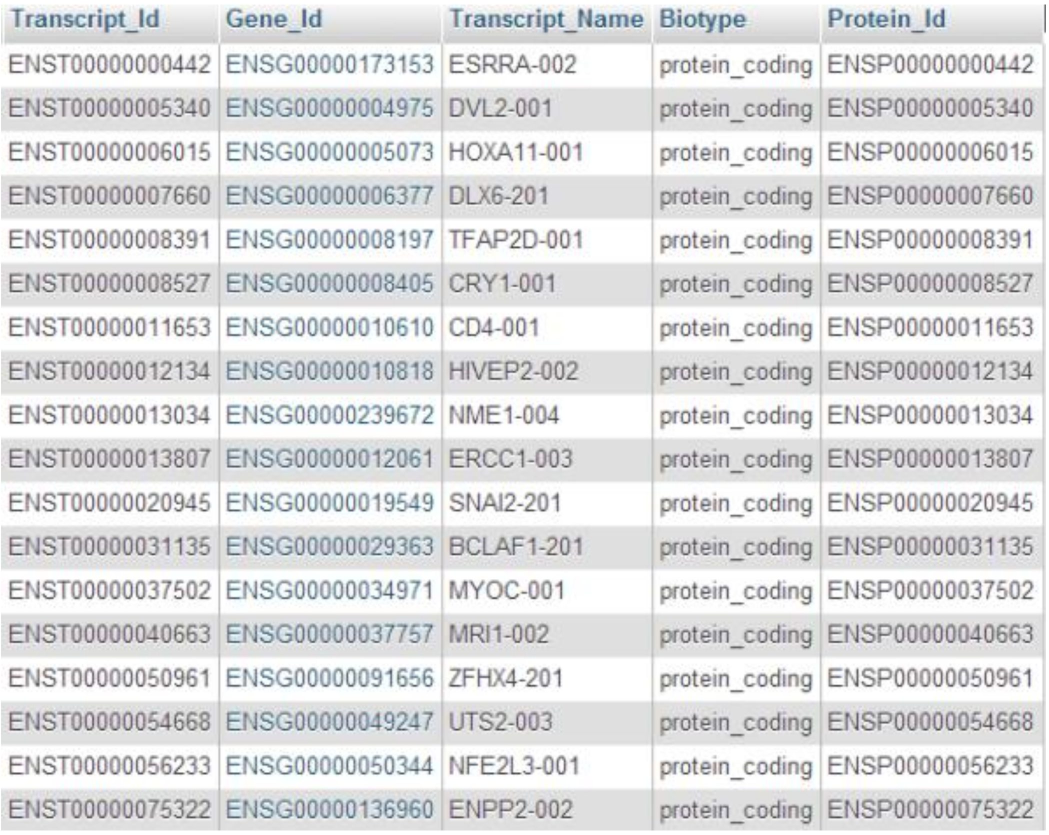
Since we need to analyze data for three species, three separate databases, one for each species namely, “Human_tr_db”, “Mouse_tr_db”, “Rat_tr_db” has been constructed.

Figures 3.11, 3.12, 3.13 and 3.14 show sample data stored in these tables for the “Human_tr_db”.



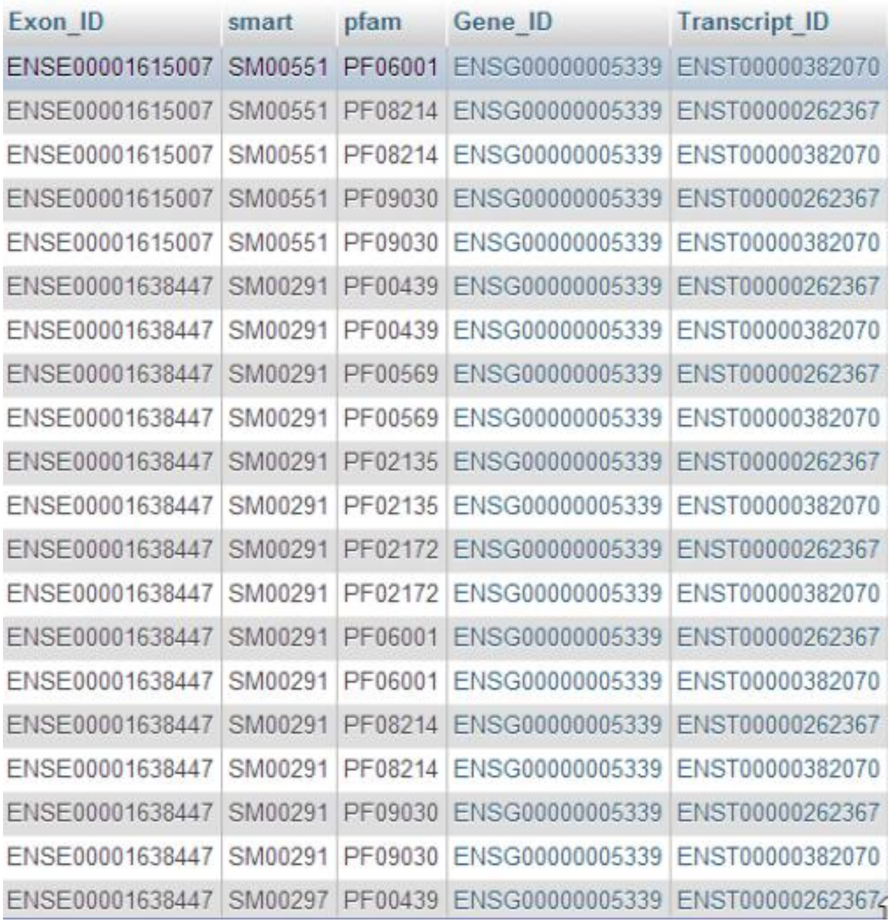
Gene_Id	Gene_Name	Description	Chr_Name
ENSG00000001084	GCLC	glutamate-cysteine ligase, catalytic subunit [Sour...	6
ENSG00000003402	CFLAR	CASP8 and FADD-like apoptosis regulator [Source:HG...	2
ENSG00000004487	KDM1A	lysine (K)-specific demethylase 1A [Source:HGNC Sy...	1
ENSG00000004848	ARX	aristaless related homeobox [Source:HGNC Symbol;Ac...	X
ENSG00000004975	DVL2	dishevelled segment polarity protein 2 [Source:HGN...	17
ENSG00000005073	HOXA11	homeobox A11 [Source:HGNC Symbol;Acc:5101]	7
ENSG00000005100	DHX33	DEAH (Asp-Glu-Ala-His) box polypeptide 33 [Source:...	17
ENSG00000005102	MEOX1	mesenchyme homeobox 1 [Source:HGNC Symbol;Acc:7013...	17
ENSG00000005302	MSL3	male-specific lethal 3 homolog (Drosophila) [Sourc...	X
ENSG00000005339	CREBBP	CREB binding protein [Source:HGNC Symbol;Acc:2348]	16
ENSG00000005436	GCFC2	GC-rich sequence DNA-binding factor 2 [Source:HGNC...	2
ENSG00000005513	SOX8	SRY (sex determining region Y)-box 8 [Source:HGNC ...	16
ENSG00000005961	ITGA2B	integrin, alpha 2b (platelet glycoprotein IIb of I...	17
ENSG00000006007	GDE1	glycerophosphodiester phosphodiesterase 1 [Source:...	16
ENSG00000006194	ZNF263	zinc finger protein 263 [Source:HGNC Symbol;Acc:13...	16
ENSG00000006377	DLX6	distal-less homeobox 6 [Source:HGNC Symbol;Acc:291...	7
ENSG00000006468	ETV1	ets variant 1 [Source:HGNC Symbol;Acc:3490]	7
ENSG00000006576	PHTF2	putative homeodomain transcription factor 2 [Sourc...	7
ENSG00000006704	GTF2IRD1	GTF2I repeat domain containing 1 [Source:HGNC Symb...	7

Figure 3.11: The "Gene_info" table sample data for "Human_tr_db"



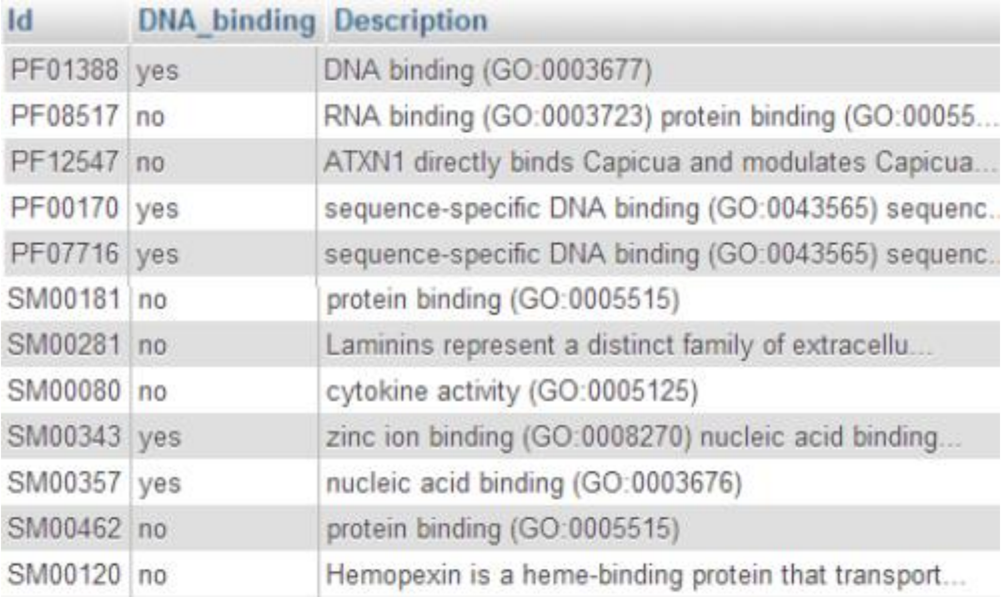
Transcript_Id	Gene_Id	Transcript_Name	Biotype	Protein_Id
ENST00000000442	ENSG00000173153	ESRRA-002	protein_coding	ENSP00000000442
ENST00000005340	ENSG00000004975	DVL2-001	protein_coding	ENSP00000005340
ENST00000006015	ENSG00000005073	HOXA11-001	protein_coding	ENSP00000006015
ENST00000007660	ENSG00000006377	DLX6-201	protein_coding	ENSP00000007660
ENST00000008391	ENSG00000008197	TFAP2D-001	protein_coding	ENSP00000008391
ENST00000008527	ENSG00000008405	CRY1-001	protein_coding	ENSP00000008527
ENST00000011653	ENSG00000010610	CD4-001	protein_coding	ENSP00000011653
ENST00000012134	ENSG00000010818	HIVEP2-002	protein_coding	ENSP00000012134
ENST00000013034	ENSG00000239672	NME1-004	protein_coding	ENSP00000013034
ENST00000013807	ENSG00000012061	ERCC1-003	protein_coding	ENSP00000013807
ENST00000020945	ENSG00000019549	SNAI2-201	protein_coding	ENSP00000020945
ENST00000031135	ENSG00000029363	BCLAF1-201	protein_coding	ENSP00000031135
ENST00000037502	ENSG00000034971	MYOC-001	protein_coding	ENSP00000037502
ENST00000040663	ENSG00000037757	MRI1-002	protein_coding	ENSP00000040663
ENST00000050961	ENSG00000091656	ZFHX4-201	protein_coding	ENSP00000050961
ENST00000054668	ENSG00000049247	UTS2-003	protein_coding	ENSP00000054668
ENST00000056233	ENSG00000050344	NFE2L3-001	protein_coding	ENSP00000056233
ENST00000075322	ENSG00000136960	ENPP2-002	protein_coding	ENSP00000075322

Figure 3.12: The "Pro_trans_info" table sample data for "Human_tr_db"



Exon_ID	smart	pfam	Gene_ID	Transcript_ID
ENSE00001615007	SM00551	PF06001	ENSG0000005339	ENST00000382070
ENSE00001615007	SM00551	PF08214	ENSG0000005339	ENST00000262367
ENSE00001615007	SM00551	PF08214	ENSG0000005339	ENST00000382070
ENSE00001615007	SM00551	PF09030	ENSG0000005339	ENST00000262367
ENSE00001615007	SM00551	PF09030	ENSG0000005339	ENST00000382070
ENSE00001638447	SM00291	PF00439	ENSG0000005339	ENST00000262367
ENSE00001638447	SM00291	PF00439	ENSG0000005339	ENST00000382070
ENSE00001638447	SM00291	PF00569	ENSG0000005339	ENST00000262367
ENSE00001638447	SM00291	PF00569	ENSG0000005339	ENST00000382070
ENSE00001638447	SM00291	PF02135	ENSG0000005339	ENST00000262367
ENSE00001638447	SM00291	PF02135	ENSG0000005339	ENST00000382070
ENSE00001638447	SM00291	PF02172	ENSG0000005339	ENST00000262367
ENSE00001638447	SM00291	PF02172	ENSG0000005339	ENST00000382070
ENSE00001638447	SM00291	PF06001	ENSG0000005339	ENST00000262367
ENSE00001638447	SM00291	PF06001	ENSG0000005339	ENST00000382070
ENSE00001638447	SM00291	PF08214	ENSG0000005339	ENST00000262367
ENSE00001638447	SM00291	PF08214	ENSG0000005339	ENST00000382070
ENSE00001638447	SM00291	PF09030	ENSG0000005339	ENST00000262367
ENSE00001638447	SM00291	PF09030	ENSG0000005339	ENST00000382070
ENSE00001638447	SM00297	PF00439	ENSG0000005339	ENST00000262367

Figure 3.13: The “Exon_Domain” table sample data for “Human_tr_db”



Id	DNA_binding	Description
PF01388	yes	DNA binding (GO:0003677)
PF08517	no	RNA binding (GO:0003723) protein binding (GO:00055...
PF12547	no	ATXN1 directly binds Capicua and modulates Capicua...
PF00170	yes	sequence-specific DNA binding (GO:0043565) sequenc...
PF07716	yes	sequence-specific DNA binding (GO:0043565) sequenc...
SM00181	no	protein binding (GO:0005515)
SM00281	no	Laminins represent a distinct family of extracellu...
SM00080	no	cytokine activity (GO:0005125)
SM00343	yes	zinc ion binding (GO:0008270) nucleic acid binding...
SM00357	yes	nucleic acid binding (GO:0003676)
SM00462	no	protein binding (GO:0005515)
SM00120	no	Hemopexin is a heme-binding protein that transport...

Figure 3.14: The “Domain_DNA_binding” table sample data for “Human_tr_db”

3.3 Hypothesis Analysis

In this study, the association between transcript diversity and protein domains in TF genes is investigated. The work done includes analysis of different human, mouse and rat RNA isoforms coded by the same TF gene, which potentially produce proteins with different domain architectures and hence functionality. The hypothesis is analyzed in several phases to follow:

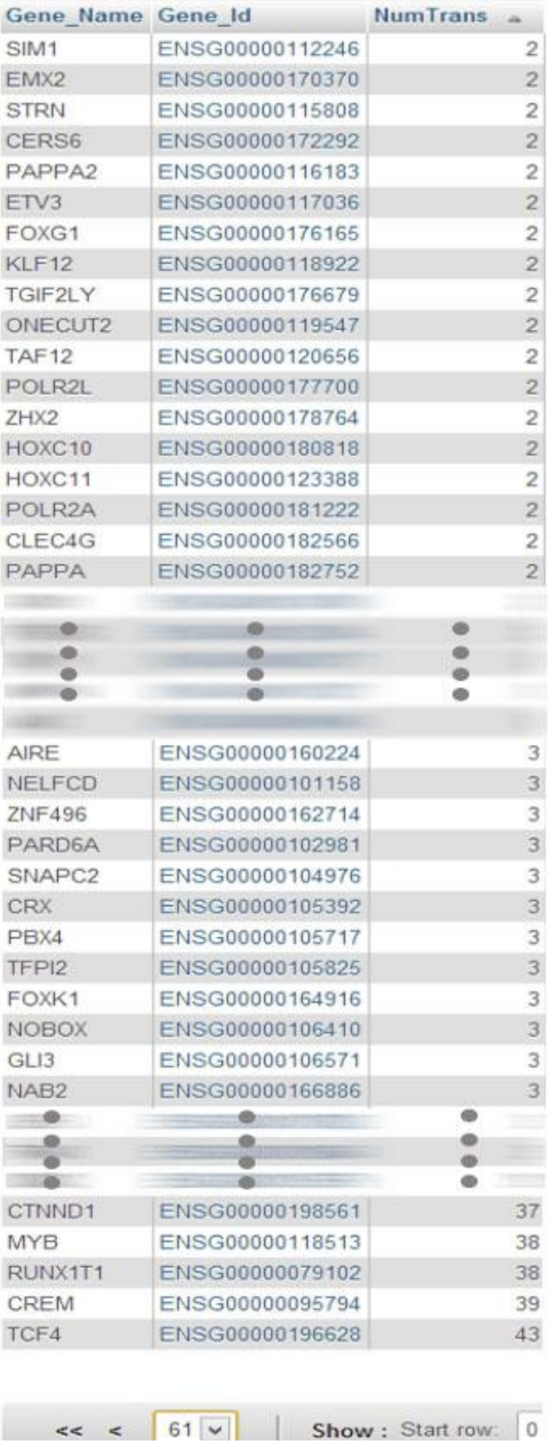
3.3.1 First Phase : Determination of TF Genes with Unique Domains

Each specific gene can produce one or more proteins. In order to determine of TF genes with unique domain firstly, TF genes with more than one transcript are found. A query is written which finds the number of TF genes with two or more transcription ids for each of the species.

In the query, first the total numbers of TF genes which produce more than one transcript are found. Using this query for each TF gene, number of transcripts is counted and the result is stored in the view referred to as "genes_with_multiple_trans". The query sent to the "Human_tr_db" database is shown in the following.

```
Select gene_info.gene_id, gene_info.gene_name
Count (pro_trans_info.transcript_id) As NumTrans
Frpm Pro_trans_info INNER JOIN gene_info
ON Gene_info.gene_id=pro_trans_info.gene_id
GROUP BY gene_info.gene_name
Having COUNT (pro_trans_info.transcript_id>1)
ORDER BY NumTrans
```

The “genes_with_multiple_trans” view has 3 columns. The first column is the TF gene name, the second column is the TF gene id and the third column is the number of transcriptions. Figure 3.15, is illustrates the schema for this view.



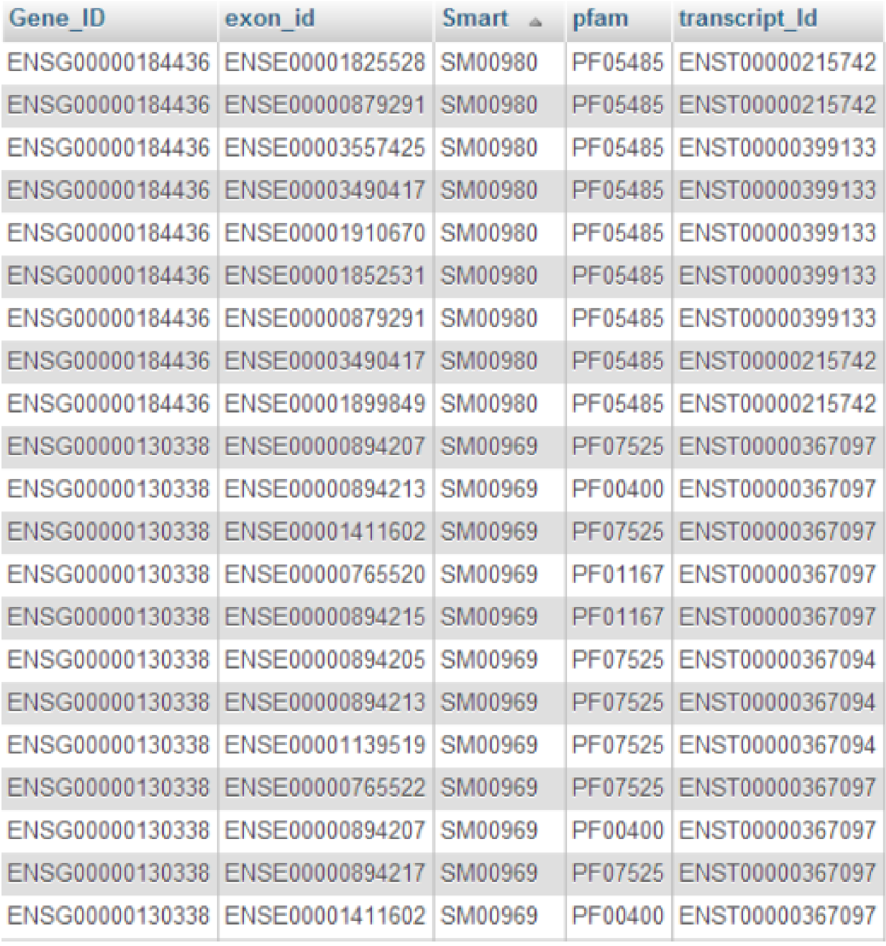
Gene_Name	Gene_Id	NumTrans
SIM1	ENSG00000112246	2
EMX2	ENSG00000170370	2
STRN	ENSG00000115808	2
CERS6	ENSG00000172292	2
PAPPA2	ENSG00000116183	2
ETV3	ENSG00000117036	2
FOXG1	ENSG00000176165	2
KLF12	ENSG00000118922	2
TGIF2LY	ENSG00000176679	2
ONECUT2	ENSG00000119547	2
TAF12	ENSG00000120656	2
POLR2L	ENSG00000177700	2
ZHX2	ENSG00000178764	2
HOXC10	ENSG00000180818	2
HOXC11	ENSG00000123388	2
POLR2A	ENSG00000181222	2
CLEC4G	ENSG00000182566	2
PAPPA	ENSG00000182752	2
⋮	⋮	⋮
AIRE	ENSG00000160224	3
NELFCD	ENSG00000101158	3
ZNF496	ENSG00000162714	3
PARD6A	ENSG00000102981	3
SNAPC2	ENSG00000104976	3
CRX	ENSG00000105392	3
PBX4	ENSG00000105717	3
TFPI2	ENSG00000105825	3
FOXK1	ENSG00000164916	3
NOBOX	ENSG00000106410	3
GLI3	ENSG00000106571	3
NAB2	ENSG00000166886	3
⋮	⋮	⋮
CTNND1	ENSG00000198561	37
MYB	ENSG00000118513	38
RUNX1T1	ENSG00000079102	38
CREM	ENSG00000095794	39
TCF4	ENSG00000196628	43

Figure 3.15: The view “genes_with_multiple_trans” for human database.

Parts of this view are joined with the “Genes_info”, “Pro_Trans_info”, “Domain_Exon” tables in the following way. For example, the rows for genes with two transcripts are extracted and joined with each of the tables mentioned. The process is repeated for parts of the view for 3 transcripts, 4 transcripts, etc. producing 43 views for human, 36 views for mouse and 8 views for rat. The procedural views are referred to as “Genes_with_(number_of_trans)_trans_info”, where (number_of_trans) is obtained as described above. These new views contain “gene_id”, “exon_id”, “transcription_id”, “SMART”, and “Pfam” attributes. The query used to produce the “Genes_with_2 trans_info“ view is shown in the following as an example.

```
CREATE VIEW genes_with_2_trans_info AS
SELECT genes_with_2_trans.gene_id, exon_domain.exon_id, Exon_domain.SMART,
Exon_doman.PFAM,Pro_trans_info.transcipt_id
FROM genes_with_2_trans
LEFT JOIN pro_trans_info ON
Genes_with_2_trans.gene_id=pro_trans_info.gene_id
LEFT JOIN Exon_domain ON
Pro_trans_info.Transcript_id=exon_domain.transcript_id
```

Figure 3.16 shows the schema for “Genes_with_2_trans” view for human database.



Gene_ID	exon_id	Smart ▲	pfam	transcript_Id
ENSG00000184436	ENSE00001825528	SM00980	PF05485	ENST00000215742
ENSG00000184436	ENSE00000879291	SM00980	PF05485	ENST00000215742
ENSG00000184436	ENSE00003557425	SM00980	PF05485	ENST00000399133
ENSG00000184436	ENSE00003490417	SM00980	PF05485	ENST00000399133
ENSG00000184436	ENSE00001910670	SM00980	PF05485	ENST00000399133
ENSG00000184436	ENSE00001852531	SM00980	PF05485	ENST00000399133
ENSG00000184436	ENSE00000879291	SM00980	PF05485	ENST00000399133
ENSG00000184436	ENSE00003490417	SM00980	PF05485	ENST00000215742
ENSG00000184436	ENSE00001899849	SM00980	PF05485	ENST00000215742
ENSG00000130338	ENSE00000894207	SM00969	PF07525	ENST00000367097
ENSG00000130338	ENSE00000894213	SM00969	PF00400	ENST00000367097
ENSG00000130338	ENSE00001411602	SM00969	PF07525	ENST00000367097
ENSG00000130338	ENSE00000765520	SM00969	PF01167	ENST00000367097
ENSG00000130338	ENSE00000894215	SM00969	PF01167	ENST00000367097
ENSG00000130338	ENSE00000894205	SM00969	PF07525	ENST00000367094
ENSG00000130338	ENSE00000894213	SM00969	PF07525	ENST00000367094
ENSG00000130338	ENSE00001139519	SM00969	PF07525	ENST00000367094
ENSG00000130338	ENSE00000765522	SM00969	PF07525	ENST00000367097
ENSG00000130338	ENSE00000894207	SM00969	PF00400	ENST00000367097
ENSG00000130338	ENSE00000894217	SM00969	PF07525	ENST00000367097
ENSG00000130338	ENSE00001411602	SM00969	PF00400	ENST00000367097

Figure 3.16: The “Genes_with_2_trans_info” view for human database.

In the first stage the protein domain diversity for each gene is analyzed by investigating the differential domain structures coded by different transcripts of the same gene. A cursor to handle a result set inside a stored procedure is defined. Domains are compared by using loops for each TF gene. If the domains are identical they are stored in the field named as “Common” otherwise they are stored in the field named as “Unique” in the view “proc_out_for(number_of_trans)_trans”, where (number_of_trans) is obtained as described above. Figure 3.17 shows the example of TF gene with two transcripts and comparison of their domains.

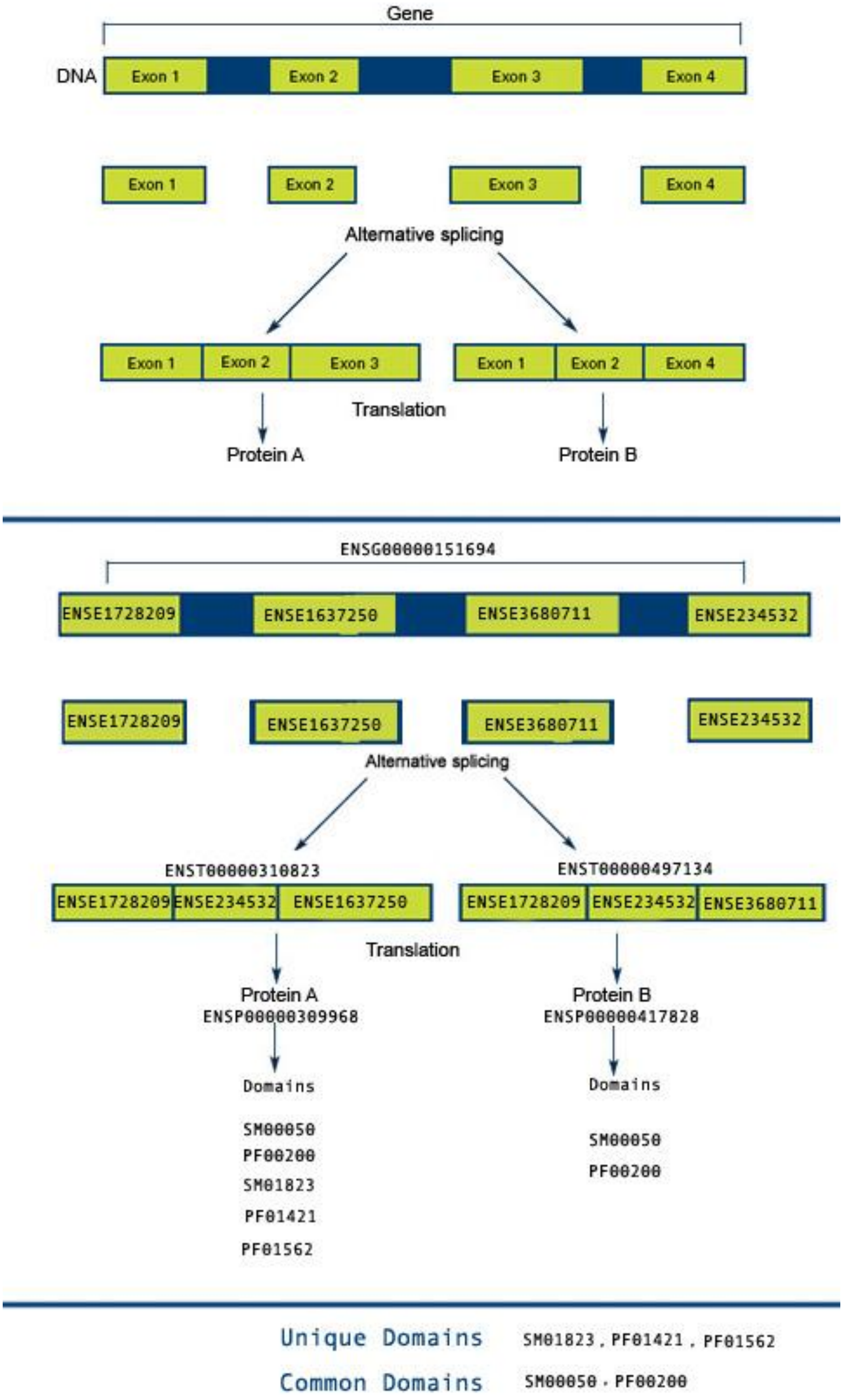


Figure 3.17: Method for analyzing protein domain diversity.

In this figure the specific gene with “ENSG00000151694” id contains four exons with “ENSE1728209”, “ENSE1637250”, “ENSE360711”, “ENSE234532” ids. With alternative splicing mechanism, two transcripts are produced from that gene, namely “ENST00000310823” and ”ENST00000497134”. Each transcript codes for one protein with ids “ENSP00000309968” and “ENSP00000417828”, respectively. The first protein contains five domains: “SM00050”, “PF00200”, “SM01823”, “PF01421” and “PF01562”. The second protein contains two domains: “SM00050” and PF00200”, as identified by SMART and PFAM databases. Comparison of the domains between these two proteins shows that both “SM00050” and PF00200” domains are common and “SM01823”, “PF01421” and “PF01562” domains are unique.

In addition, for each gene, the total number of transcripts available are compared with one another, and unique domains that are present in only one transcript are reported.

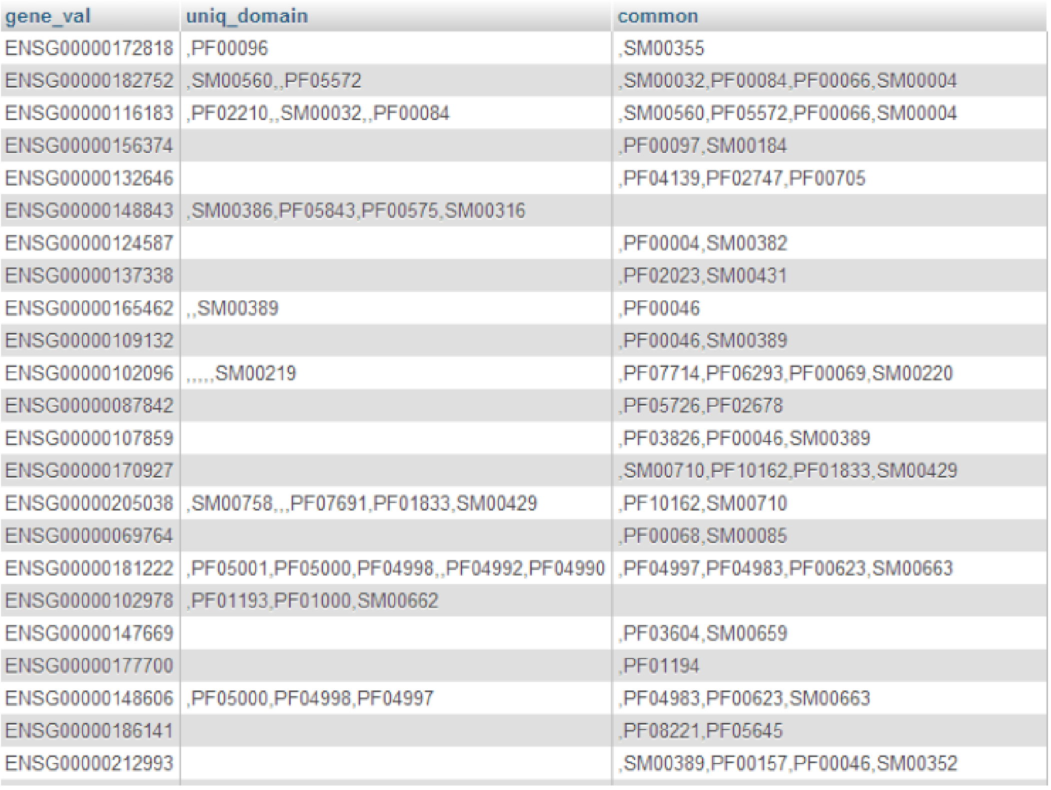
The sample code for procedure “proc_out_for_2_trans” follows:

```
begin
 -- Variables Declaration;
-- Cursor Declaration;

  DECLARE gene2trans CURSOR FOR

    SELECT

        Gene_Id,Transcript_Id,smart,pfam

    FROM genes_with_2_trans_info

    WHERE Gene_Id = gene_in;

  -- 'handlers' for exceptions Declaration
DECLARE CONTINUE HANDLER FOR NOT FOUND

    SET no_more_rows = TRUE;
```

```
OPEN gene2trans;

 Select FOUND_ROWS() into num_rows;

 the_loop: LOOP

              FETCH gene2trans
 INTO   gene_val,

        transcript_val,sm,pf;

    IF no_more_rows THEN

        CLOSE gene2trans;

        LEAVE the_loop;

    END IF;

            set num_trans=num_trans+1;

            set first_transcript_val= transcript_val;

            while  first_transcript_val like transcript_val do

                        if instr(uniq1,sm)=0 then

                                    select concat(sm,',',uniq1) into uniq1;

                        end if;

                        if instr(uniq1,pf)=0 then

                                    select concat(pf,',',uniq1) into uniq1;

                        end if;

                         FETCH  gene2trans

            INTO   gene_val,

        transcript_val,sm,pf;

            IF no_more_rows THEN

                         CLOSE gene2trans;

                         LEAVE the_loop;

                        END IF;
            if first_transcript_val not like transcript_val then
                set num_trans=num_trans+1;
                set trans_one=first_transcript_val;
                set first_transcript_val= transcript_val;
                set set_one=uniq1;
                set uniq1='';
            end if;
        end while;
        SET loop_cntr = loop_cntr + 1;
END LOOP the_loop;
 select trans_one as first_transcript,set_one as first_set,transcript_val as
second_transcript, uniq1 as second_set,common,uniq_domain,uniq_trans;

while uniq1<> '' do
        select locate(',',uniq1)into pos;
        select substr(uniq1,1,(pos-1))into res;
        select substr(uniq1,pos+1,leng1-pos)into uniq1;
        if (locate(res,set_one)<> 0) then
            select concat(common,',',res) into common;
            select replace(set_one,res,'') into set_one;
        elseif  (locate(res,set_one)= 0)then
            select concat(uniq_domain,',',res) into uniq_domain;
            set uniq_trans=transcript_val;
        end if;

end while;
```

```
select trim(','from set_one ) into set_one;

 if length(set_one)>2  then

select concat(uniq_trans,',',trans_one ) into uniq_trans;

end if;

 select concat(uniq_domain,',',set_one) into uniq_domain;

select trans_one as first_transcript,set_one as first_set,transcript_val as
second_transcript, uniq1 as second_set ,common,uniq_domain,uniq_trans;

end
```

The view constructed after this step is illustrated in Figure 3.18. The results are analyzed and discussed in Chapter 4.



gene_val	uniq_domain	common
ENSG00000172818	,PF00096	,SM00355
ENSG00000182752	,SM00560,,PF05572	,SM00032,PF00084,PF00066,SM00004
ENSG00000116183	,PF02210,,SM00032,,PF00084	,SM00560,PF05572,PF00066,SM00004
ENSG00000156374		,PF00097,SM00184
ENSG00000132646		,PF04139,PF02747,PF00705
ENSG00000148843	,SM00386,PF05843,PF00575,SM00316	
ENSG00000124587		,PF00004,SM00382
ENSG00000137338		,PF02023,SM00431
ENSG00000165462	,,SM00389	,PF00046
ENSG00000109132		,PF00046,SM00389
ENSG00000102096	,,,,,SM00219	,PF07714,PF06293,PF00069,SM00220
ENSG00000087842		,PF05726,PF02678
ENSG00000107859		,PF03826,PF00046,SM00389
ENSG00000170927		,SM00710,PF10162,PF01833,SM00429
ENSG00000205038	,SM00758,,,PF07691,PF01833,SM00429	,PF10162,SM00710
ENSG00000069764		,PF00068,SM00085
ENSG00000181222	,PF05001,PF05000,PF04998,,PF04992,PF04990	,PF04997,PF04983,PF00623,SM00663
ENSG00000102978	,PF01193,PF01000,SM00662	
ENSG00000147669		,PF03604,SM00659
ENSG00000177700		,PF01194
ENSG00000148606	,PF05000,PF04998,PF04997	,PF04983,PF00623,SM00663
ENSG00000186141		,PF08221,PF05645
ENSG00000212993		,SM00389,PF00157,PF00046,SM00352

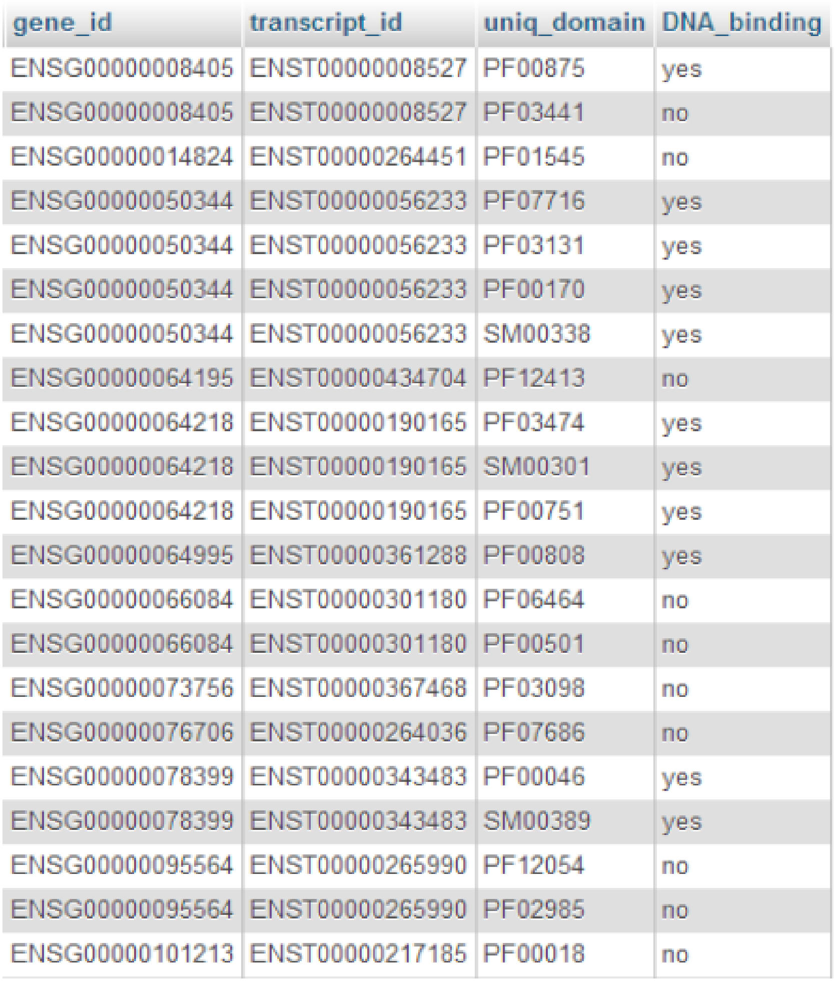
Figure 3.18: TF genes with unique and common domains in human database.

3.3.2 Second Phase : Domains with DNA-Binding Function

The aim of this phase is to determine the TF genes with 2 transcripts with unique domains, which illustrate differential DNA binding ability. In order to achieve this goal, domains, which have DNA binding ability in SMART, and Pfam databases must be identified.

Firstly, the concept for “DNA binding” property is searched for unique domains of TF genes that produce two transcripts. The phrases “DNA-binding”, “DNA binding activity”, “bind to DNA”, “Nucleic Acid binding”, ”chromatin binding” are used to search for this property in SMART and Pfam databases.

A new table named “Domains_DNA_binding” which contains the domains and notation about their ability to bind DNA is added to each of the three databases for each species. In order to analyze the TF gene which has DNA-binding ability, initially a query is written which joins the output of the procedure in the previous phase with the “Domains_DNA_binding” table. The output of this join is stored in a new view referred to as “DNA_Binding”. Figure 3.19 illustrates the “DNA_Binding” view.



gene_id	transcript_id	uniq_domain	DNA_binding
ENSG00000008405	ENST00000008527	PF00875	yes
ENSG00000008405	ENST00000008527	PF03441	no
ENSG00000014824	ENST00000264451	PF01545	no
ENSG00000050344	ENST00000056233	PF07716	yes
ENSG00000050344	ENST00000056233	PF03131	yes
ENSG00000050344	ENST00000056233	PF00170	yes
ENSG00000050344	ENST00000056233	SM00338	yes
ENSG00000064195	ENST00000434704	PF12413	no
ENSG00000064218	ENST00000190165	PF03474	yes
ENSG00000064218	ENST00000190165	SM00301	yes
ENSG00000064218	ENST00000190165	PF00751	yes
ENSG00000064995	ENST00000361288	PF00808	yes
ENSG00000066084	ENST00000301180	PF06464	no
ENSG00000066084	ENST00000301180	PF00501	no
ENSG00000073756	ENST00000367468	PF03098	no
ENSG00000076706	ENST00000264036	PF07686	no
ENSG00000078399	ENST00000343483	PF00046	yes
ENSG00000078399	ENST00000343483	SM00389	yes
ENSG00000095564	ENST00000265990	PF12054	no
ENSG00000095564	ENST00000265990	PF02985	no
ENSG00000101213	ENST00000217185	PF00018	no

Figure 3.19: “DNA_Binding” view for human database.

Then, by using a query as shown in the following the SMART and Pfam domains, which have DNA-binding ability, are counted separately for each species. Results are presented in Chapter 4.

SELECT COUNT (gene_id)

FROM DNA_binding_domain

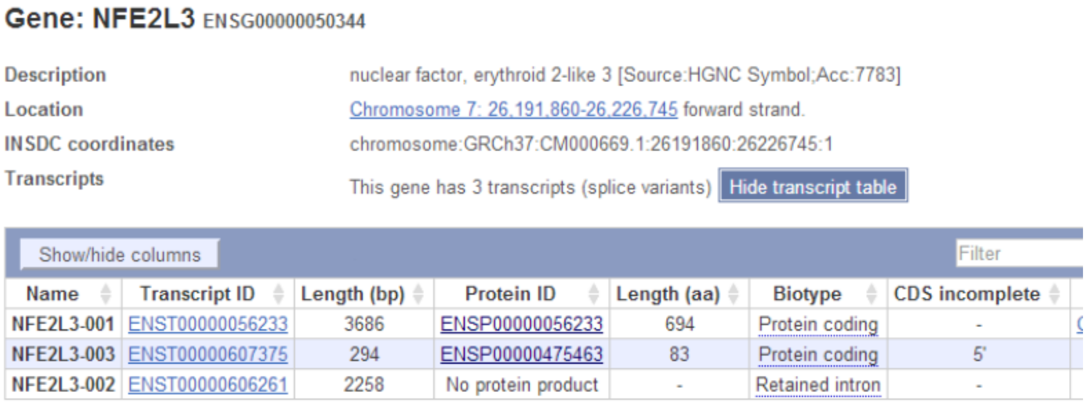
WHERE DNA_binding=’YES’

AND unique_domain LIKE “SM%”

Finally, another query is written in order to count the number of TF genes that have DNA-binding ability. The result is presented in Chapter 4. The query is used for this purpose is shown in the following.

```
SELECT COUNT (DISTINCT gene_id)
FROM DNA_binding_domain
WHERE DNA_binding='YES'
```

For example, the NFE213 TF gene with "ENSG00000050344" id, shown in Figure 3.20 has three transcripts with "ENST00000056233", "ENST00000607375" and "ENST00000606261" ids, but only two of them ("ENST00000056233", "ENST00000607375") produce proteins with "ENSP00000056233" and "ENSP0000047475463" ids.



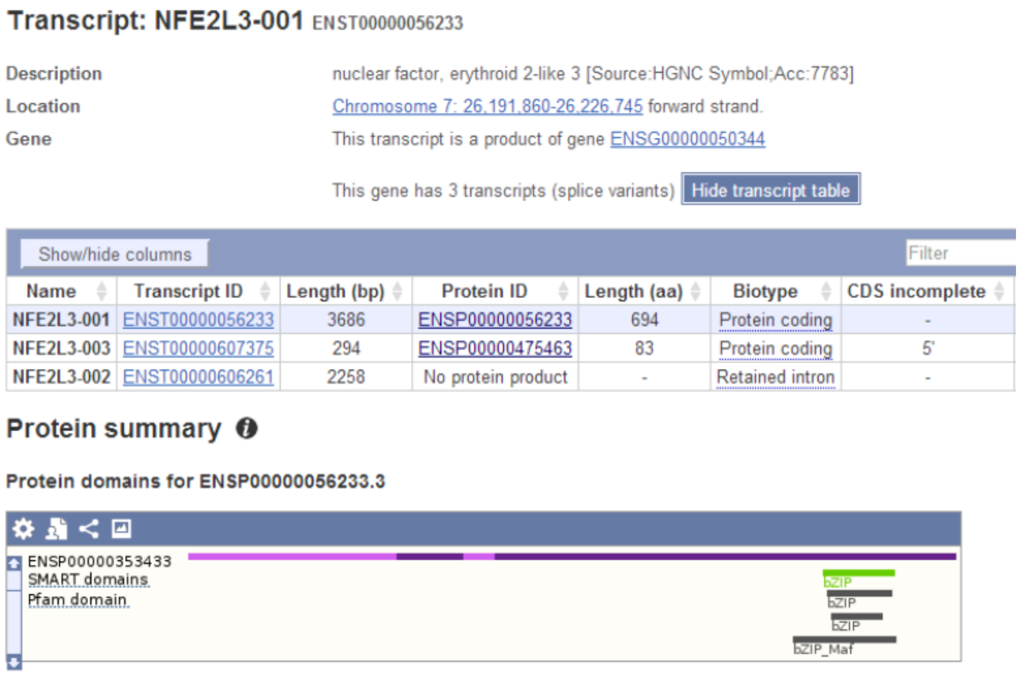
Gene: NFE2L3 ENSG00000050344

Description: nuclear factor, erythroid 2-like 3 [Source:HGNC Symbol;Acc:7783]
Location: Chromosome 7: 26,191,860-26,226,745 forward strand.
INSDC coordinates: chromosome:GRCh37:CM000669.1:26191860:26226745:1
Transcripts: This gene has 3 transcripts (splice variants)

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CDS incomplete
NFE2L3-001	ENST00000056233	3686	ENSP00000056233	694	Protein coding	-
NFE2L3-003	ENST00000607375	294	ENSP00000475463	83	Protein coding	5'
NFE2L3-002	ENST00000606261	2258	No protein product	-	Retained intron	-

Figure 3.20: The NFE213 TF gene information from Ensembl. (The figure is taken from [59]).

Figure 3.21 shows the protein sequence and domains for "ENSP00000056233" in protein summary part. This protein contains one SMART and three Pfam domains.



Transcript: NFE2L3-001 ENST00000056233

Description: nuclear factor, erythroid 2-like 3 [Source:HGNC Symbol;Acc:7783]

Location: Chromosome 7: 26,191,860-26,226,745 forward strand.

Gene: This transcript is a product of gene ENSG00000050344

This gene has 3 transcripts (splice variants) Hide transcript table

Show/hide columns Filter

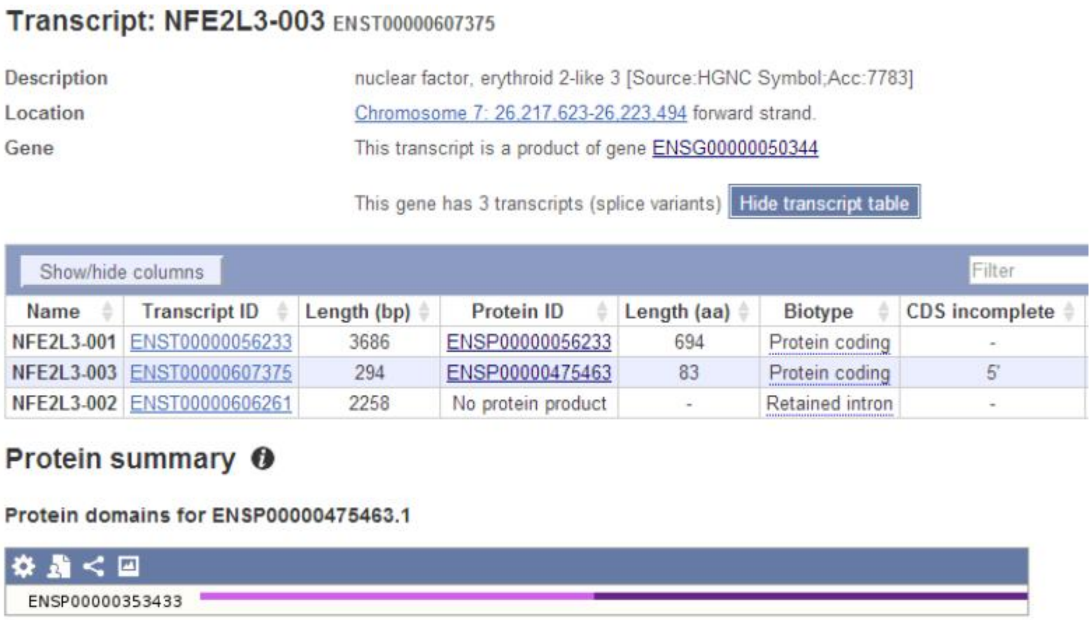
Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CDS incomplete
NFE2L3-001	ENST00000056233	3686	ENSP00000056233	694	Protein coding	-
NFE2L3-003	ENST00000607375	294	ENSP00000475463	83	Protein coding	5'
NFE2L3-002	ENST00000606261	2258	No protein product	-	Retained intron	-

Protein summary

Protein domains for ENSP00000056233.3

Figure 3.21: Protein domains information for “ENSP00000056233”. (The figure is taken from [59]).

Figure 3.22 shows that “ENSP00000475463” does not have any domains.



Transcript: NFE2L3-003 ENST00000607375

Description: nuclear factor, erythroid 2-like 3 [Source:HGNC Symbol;Acc:7783]

Location: Chromosome 7: 26,217,623-26,223,494 forward strand.

Gene: This transcript is a product of gene ENSG00000050344

This gene has 3 transcripts (splice variants) Hide transcript table

Show/hide columns Filter

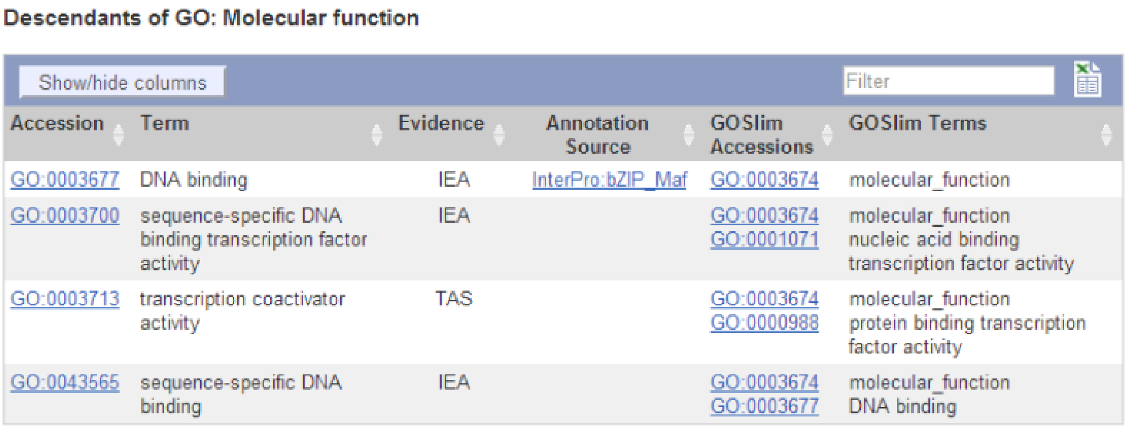
Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CDS incomplete
NFE2L3-001	ENST00000056233	3686	ENSP00000056233	694	Protein coding	-
NFE2L3-003	ENST00000607375	294	ENSP00000475463	83	Protein coding	5'
NFE2L3-002	ENST00000606261	2258	No protein product	-	Retained intron	-

Protein summary

Protein domains for ENSP00000475463.1

Figure 3.22: The “ENSP00000475463” with no domain. (The figure is taken from [59]).

Further analysis depicted in Figure 3.23 shows that the Pfam domain named "bZIP_Maf" has DNA-binding ability. So, this protein should be counted as a protein with DNA-binding function. Figure 3.23 shows the Gene Ontology (GO) molecular function of this domain.



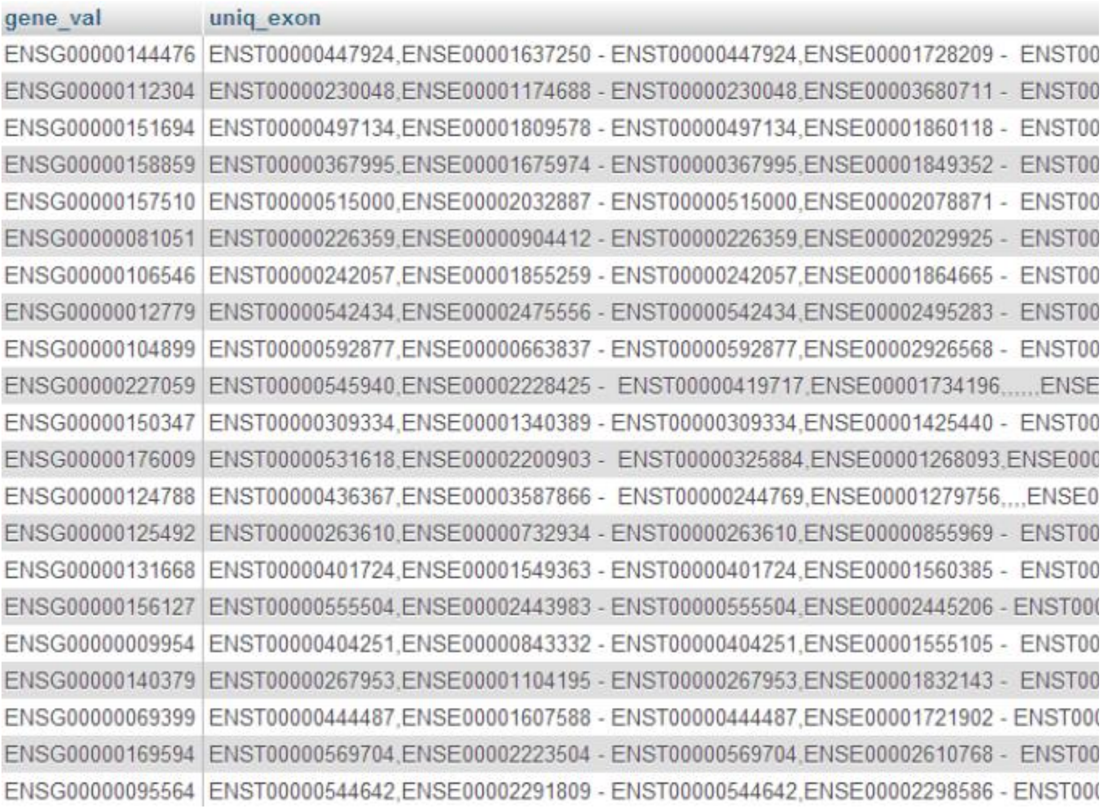
Descendants of GO: Molecular function

Accession	Term	Evidence	Annotation Source	GOSlim Accessions	GOSlim Terms
GO:0003677	DNA binding	IEA	InterPro:bZIP_Maf	GO:0003674	molecular_function
GO:0003700	sequence-specific DNA binding transcription factor activity	IEA		GO:0003674 GO:0001071	molecular_function nucleic acid binding transcription factor activity
GO:0003713	transcription coactivator activity	TAS		GO:0003674 GO:0000988	molecular_function protein binding transcription factor activity
GO:0043565	sequence-specific DNA binding	IEA		GO:0003674 GO:0003677	molecular_function DNA binding

Figure 3.23: Example of domains with and without DNA- binding ability. (The figure is taken from [59]).

3.3.3 Third Phase : Determination of TF Genes with Unique Exons

The unique exon ids for TF genes which have two transcripts are determined in this phase. First of all, some comparison methods are used. For each TF gene common and unique exons are found. A procedure is used for this purpose. Determination of the transcripts to which each exon belongs to it significant. Figure 3.24 shows the output of the procedure.



gene_val	uniq_exon
ENSG00000144476	ENST00000447924,ENSE00001637250 - ENST00000447924,ENSE00001728209 - ENST00
ENSG00000112304	ENST00000230048,ENSE00001174688 - ENST00000230048,ENSE00003680711 - ENST00
ENSG00000151694	ENST00000497134,ENSE00001809578 - ENST00000497134,ENSE00001860118 - ENST00
ENSG00000158859	ENST00000367995,ENSE00001675974 - ENST00000367995,ENSE00001849352 - ENST00
ENSG00000157510	ENST00000515000,ENSE00002032887 - ENST00000515000,ENSE00002078871 - ENST00
ENSG00000081051	ENST00000226359,ENSE00000904412 - ENST00000226359,ENSE00002029925 - ENST00
ENSG00000106546	ENST00000242057,ENSE00001855259 - ENST00000242057,ENSE00001864665 - ENST00
ENSG00000012779	ENST00000542434,ENSE00002475556 - ENST00000542434,ENSE00002495283 - ENST00
ENSG00000104899	ENST00000592877,ENSE00000663837 - ENST00000592877,ENSE00002926568 - ENST00
ENSG00000227059	ENST00000545940,ENSE00002228425 - ENST00000419717,ENSE00001734196,,,,,,ENSE
ENSG00000150347	ENST00000309334,ENSE00001340389 - ENST00000309334,ENSE00001425440 - ENST00
ENSG00000176009	ENST00000531618,ENSE00002200903 - ENST00000325884,ENSE00001268093,ENSE000
ENSG00000124788	ENST00000436367,ENSE00003587866 - ENST00000244769,ENSE00001279756,,,,ENSE0
ENSG00000125492	ENST00000263610,ENSE00000732934 - ENST00000263610,ENSE00000855969 - ENST00
ENSG00000131668	ENST00000401724,ENSE00001549363 - ENST00000401724,ENSE00001560385 - ENST00
ENSG00000156127	ENST00000555504,ENSE00002443983 - ENST00000555504,ENSE00002445206 - ENST00
ENSG00000009954	ENST00000404251,ENSE00000843332 - ENST00000404251,ENSE00001555105 - ENST00
ENSG00000140379	ENST00000267953,ENSE00001104195 - ENST00000267953,ENSE00001832143 - ENST00
ENSG00000069399	ENST00000444487,ENSE00001607588 - ENST00000444487,ENSE00001721902 - ENST00
ENSG00000169594	ENST00000569704,ENSE00002223504 - ENST00000569704,ENSE00002610768 - ENST00
ENSG00000095564	ENST00000544642,ENSE00002291809 - ENST00000544642,ENSE00002298586 - ENST00

Figure 3.24: TF genes with unique exons.

3.3.4 Statististical Analysis

Following statistical analysis were performed to evaluate whether there are any significant differences of results across species:

Table 3.3: Species.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Human	106	50.0	50.0	50.0
	Mouse	87	41.0	41.0	91.0
	Rat	19	9.0	9.0	100.0
	Total	212	100.0	100.0	

Table 3.4: Species number of TF genes with DNA binding ability Crosstabulation.

			Number of TF genes with DNA binding ability		Total
			Not Present	Present	
Species	Human	Count	52	54	106
		Expected Count	53.0	53.0	106.0
		% within Species	49.1%	50.9%	100.0%
		% of Total	24.5%	25.5%	50.0%
		Std. Residual	-.1	.1	
	Mouse	Count	42	45	87
		Expected Count	43.5	43.5	87.0
		% within Species	48.3%	51.7%	100.0%
		% of Total	19.8%	21.2%	41.0%
		Std. Residual	-.2	.2	
	Rat	Count	12	7	19
		Expected Count	9.5	9.5	19.0
		% within Species	63.2%	36.8%	100.0%
		% of Total	5.7%	3.3%	9.0%
		Std. Residual	.8	-.8	
Total		Count	106	106	212
		Expected Count	106.0	106.0	212.0
		% within Species	50.0%	50.0%	100.0%
		% of Total	50.0%	50.0%	100.0%

Table 3.5:Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.457[a]	2	.483
Likelihood Ratio	1.473	2	.479
Linear-by-Linear Association	.546	1	.460
N of Valid Cases	212		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 9.50.

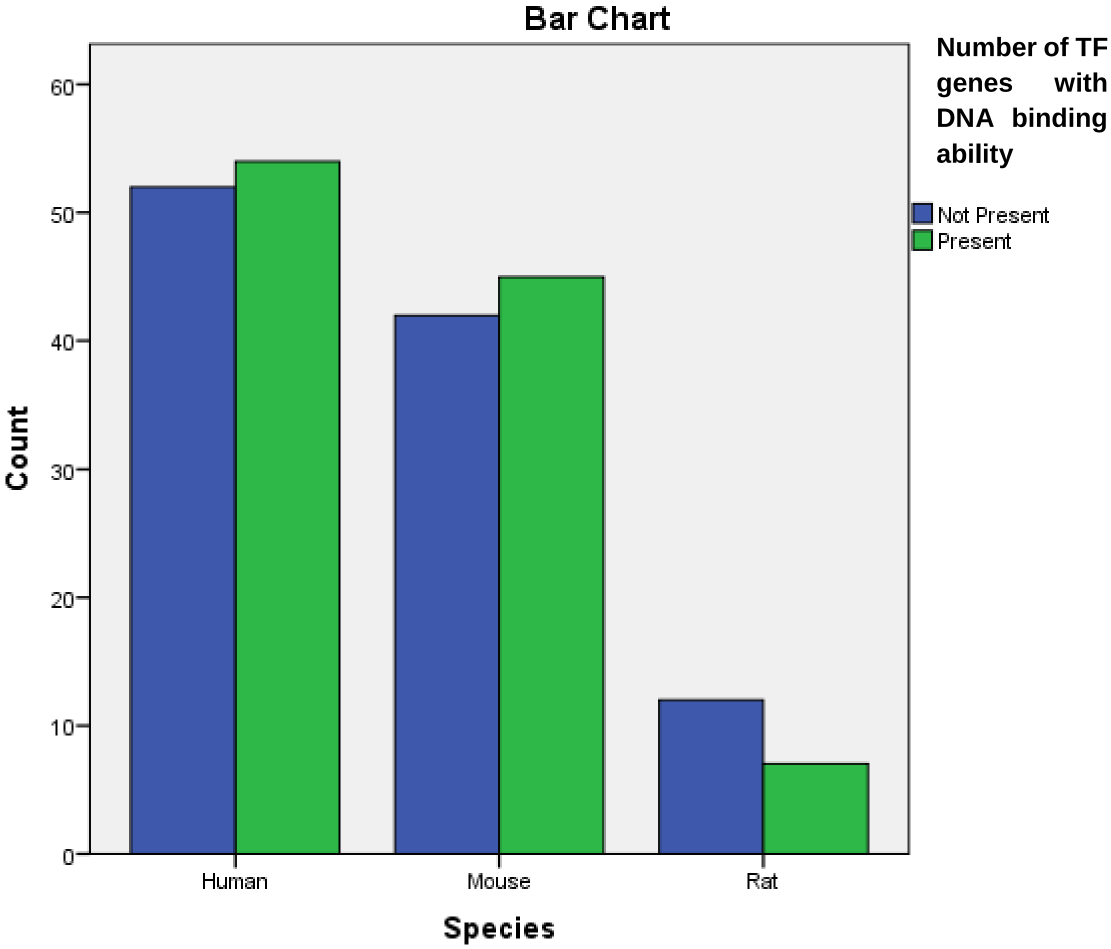


Figure 3.25: Distribution of TF genes with DNA binding ability.

The null hypothesis was tested using Chi-squared test to see whether there was a difference between the percentage of TF genes with unique domains that show DNA binding by species. The percentage of TF genes with unique domains that show DNA binding did not differ by species,

χ^2 (2, N = 212) = 1.457, p = 0.483, χ^2(1, N = 90) = 0.89, p = 0.35.

Chapter 4

RESULTS AND DISCUSSION

In this thesis, the association between transcript diversity and protein domains for three eukaryotic species is studied. The analysis is conducted for human, mouse and rat. There are a lot of sequence data in these three species to study. Firstly, the number of TF genes for each species is counted using the NCBI database. Table 4.1 shows the number of TF genes for each species. It can be seen that human genome possesses the highest number of TF genes.

Table 4.1: Total number of TF genes in each genome.

Species	Number of TF genes
Human	2152
Mouse	1567
Rat	1150

Next, the TF genes are categorized according to the total number of transcripts available for each gene in NCBI. This data is shown in Table 4.2. As evident from this table, human TF genes with multiple transcript sequences are higher than those in mouse and the ones in mouse are higher than those in rat. The reason for these differences could be in the fact that human TF transcripts are more widely studied and sequenced compared to the other two model organisms. Similarly, mouse presents more data, as it is a more commonly studied species.

Table 4.2: The number of transcripts for three species.

Number of transcripts per gene	Number of human TF genes	Number of mouse TF genes	Number of rat TF genes
1	346	622	1033
2	306	335	94
3	269	182	14
4	242	135	4
5	199	97	3
6	136	60	1
7	108	38	0
8	100	24	1
9	91	19	0
10	66	12	0
11	57	9	0
12	55	10	0
13	31	4	0
14	25	5	0
15	26	4	0
16	14	3	0
17	23	4	0
18	12	1	0
19	9	1	0
20	8	0	0
21	4	0	0
22	4	1	0
23	1	0	0
24	6	0	0
27	2	0	0
28	3	0	0
29	1	0	0
30	2	0	0
31	1	0	0
36	0	1	0
37	1	0	0
38	2	0	0
39	1	0	0
43	1	0	0
Total	2152	1567	1150

Genes with a single transcript or multiple number of transcripts are separated into two main categories. The TF genes which produce more than one transcript is of interest for further analysis in this thesis. In the following sections, further detailed analysis on such genes with multiple transcripts within each of the three genomes are presented.

4.1 Human TF Transcript Analysis

In the human genome, a total of 2152 TF genes are analyzed. Figure 4.1 shows the distribution of total number of transcripts per gene. In this figure, x-axis represents the total number of transcripts sequenced per gene. The y-axis represents the total number of TF genes. As evident from this figure, the majority of human TF genes have 1 or 2 transcripts. As the number of transcripts sequenced per gene increases, the total number of genes decreases.

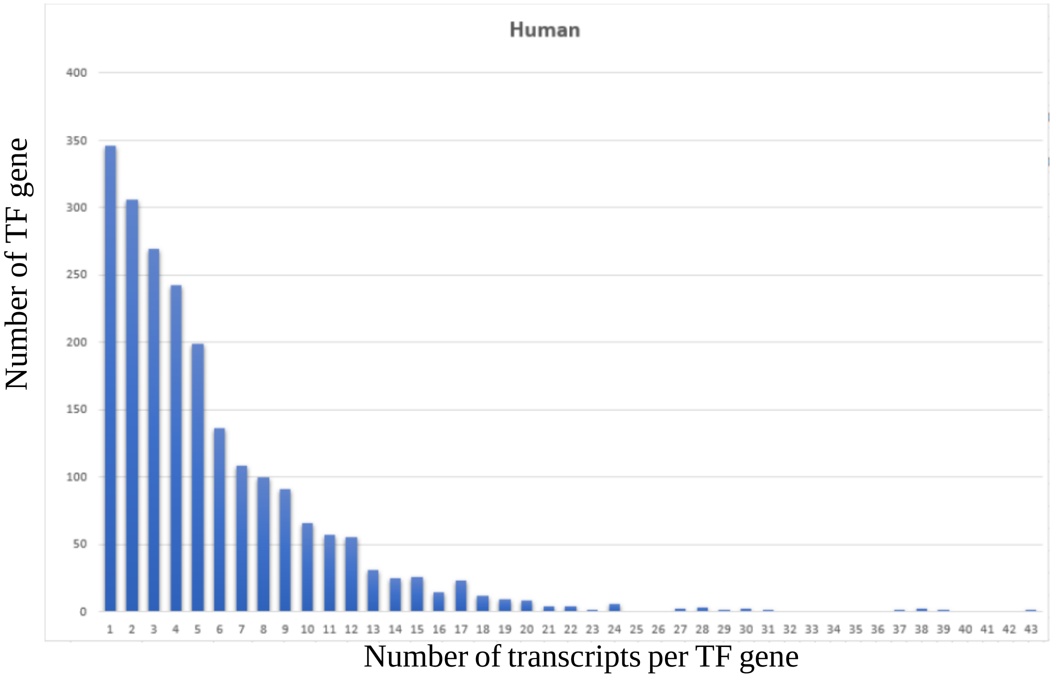


Figure 4.1: Distribution of TF genes with different numbers of transcripts in human.

4.1.1 TF Gene Categories

Table 4.3 shows the number of TF genes in human species that produce one or more transcripts. It can be seen that the majority (83.92%) of TF genes have multiple transcripts.

Table 4.3: Distribution of TF genes for human.

Number of TF genes with	Number	Percentage
one transcript only	346	16.08%
multiple (two or more) transcripts	1806	83.92%

Of the genes with multiple transcripts, the genes that have 2 transcripts are analyzed in terms of their protein coding and their protein structures. This is further explained in the following sections.

4.1.2 Number of TF Genes With Unique Domains

Table 4.4 shows the number of TF genes with multiple number of transcripts for human. In addition, this table provides an analysis for the protein domains coded by these transcripts. For each gene, the total number of transcripts available are compared with one another, and unique domains that are present in only one transcript and not the other are reported. As shown in Table 4.4, many genes have transcripts with unique domains. For example, this is the case for 35% of genes with 2 transcripts only and 21% of genes with 3 transcripts only. It would be expected that the percentage of genes with unique domains would increase as the transcript numbers increase. However, this is not the case. The reason for this is because most of the genes with higher number of transcripts, actually have short sequences such as ESTs, and not full-length transcripts.

Table 4.4: Number of human TF genes with unique domains that is present in only one transcript.

Number of Transcripts per TF gene	Number of TF genes	Number of TF genes with unique domains	% of TF genes with unique domain
2	306	106	34.64%
3	269	57	21.18%
4	242	48	19.83%
5	199	27	13.56%
6	136	13	9.56%
7	108	8	7.41%
8	100	11	11%
9	91	5	5.49%
10	66	5	7.57%
11	57	3	5.26%
12	55	1	1.81%
13	31	0	0
14	25	0	0
15	26	2	7.69%
16	14	1	7.14%
17	23	1	4.76%
18	12	0	0
19	9	2	22.2%
20	8	0	0
21	4	0	0
22	4	0	0
23	1	0	0
24	6	1	16.6%
27	2	0	0
28	3	1	33.3%
29	1	0	0
30	2	0	0
31	1	0	0
37	1	0	0
38	2	0	0
39	1	0	0
43	1	0	0

4.1.3 Domains with DNA-Binding Function

TFs are proteins that regulate transcription. Each TF binds to one particular set of DNA sequence. For this reason TFs include DNA-binding domain, which modulate the process of transcription. The role of DNA binding domain is to bring the transcription activation domain into the vicinity of the preintiation complex of transcription.

In this study, further analyses of transcript and domain relationships are restricted to human TF genes with 2 transcripts only. Such genes are in total 306. A total of 106 of these genes have transcripts with unique domains. Further analysis focus on these 106 genes. The transcripts coded by these 106 genes bring in total of 247 unique domains. Table 4.5 shows further functional analysis on these domains.

These genes are analyzed in terms of the functions of their unique domains. In particular, DNA binding properties of these domains are investigates.

The total number of domains reported in SMART and Pfam for TF genes which produce 2 transcripts with unique domains and the number of unique domains with DNA binding ability is shown in Table 4.5. The percentage of such TF genes is also shown.

Table 4.5: Number and percentage of domains with DNA binding ability from human TF genes with 2 transcripts.

Database	Total number of unique domains	Number of domains with DNA binding ability	Percentage
Pfam	151	68	45.03%
SMART	96	42	43.75%

As evident from Table 4.5, a substantial portion of unique domains analyzed within Pfam and SMART have DNA binding ability.

Lastly, the genes coding for these unique domains are analyzed. Number of TF genes with 2 transcripts, which have DNA-binding ability accounts for about 51% of the genes analyzed in this category, as shown in Table 4.6. This number is important since it shows that more than 50 percent of proteins, which are products of TF genes with 2 transcripts, have DNA binding ability. This result indicates that at least half of the unique domains introduced by different transcripts of a human TF gene deliver DNA binding ability.

Table 4.6: Number and percentage of human TF genes with 2 transcripts which have DNA-binding ability.

Total number of genes with unique domains	Number of TF genes with unique domains that show DNA binding ability	Percentage
106	54	50.94%

4.2 Mouse TF Transcript Analysis

In the mouse genome, a total of 1567 TF genes are analyzed. Figure 4.2 shows the distribution of total number of transcripts per gene. In this figure, x-axis represents the total number of transcripts sequenced per gene. The y-axis represents the total number of TF genes. As evident from this figure, the majority of mouse TF genes have 1 or 2 transcripts. As the number of transcripts sequenced per gene increase, the total number of genes decrease.

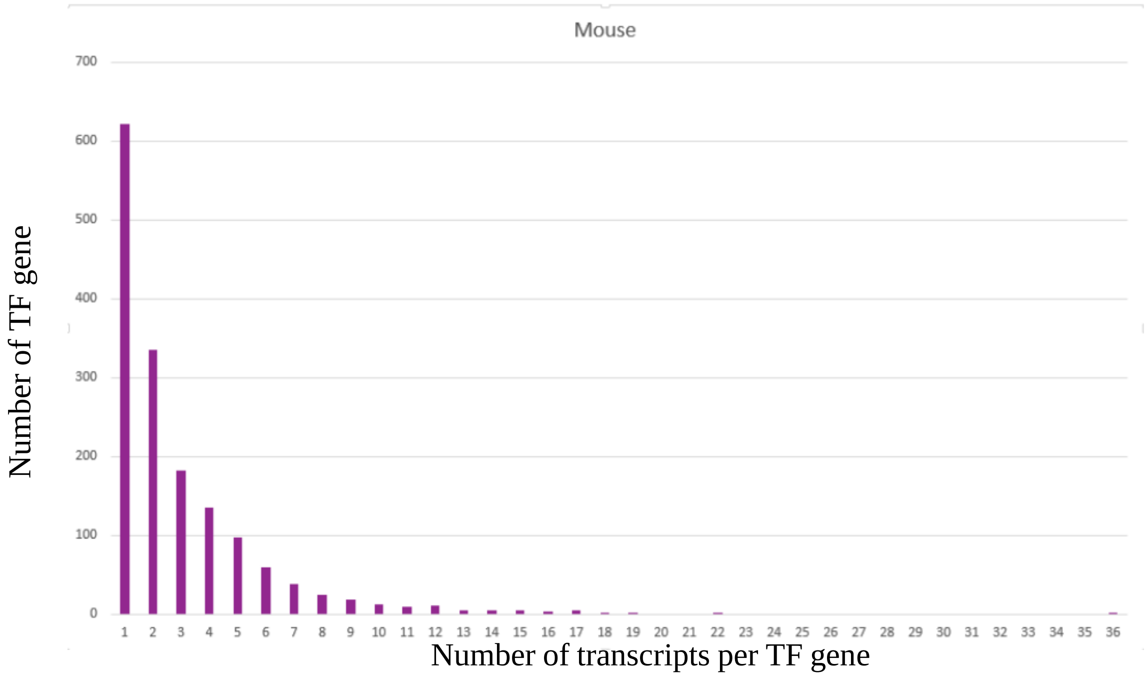


Figure 4.2: Distribution of TF genes with different numbers of transcripts in mouse.

4.2.1 TF Gene Categories

Table 4.7 shows the number of TF genes in mouse species that produce one or more transcripts. It can be seen that the majority (60.31%) of TF genes have multiple transcripts.

Table 4.7: Distribution of TF genes for mouse.

Number of TF genes with	Number	Percentage
One transcript only	622	39.69%
Multiple (two or more) transcripts	945	60.31%

Of the genes with multiple transcripts, the genes that have 2 transcripts are analyzed in terms of their protein coding and their protein structures. This is further explained in the following sections.

4.2.2 Number of TF Genes With Unique Domains

Table 4.8 shows the number of TF genes with multiple different number of transcripts for mouse. In addition, this table provides an analysis for the protein domains coded by these transcripts. For each gene, the total number of transcripts available are compared with one another, and unique domains that are present in only one transcript and not the other are reported. As shown in Table 4.8, many genes have transcripts with unique domains. For example, this is the case for 26% of genes with 2 transcripts only and 21% of genes with 3 transcripts only. It would be expected that the percent genes with unique domains would increase as the transcript numbers increase. However, this is not the case. The reason for this is because most of the genes with higher number of transcripts, actually have short sequences such as ESTs, and not full-length transcripts.

Table 4.8: Number of mouse TF genes with unique domains that is present in only one transcript.

Number of transcripts per TF gene	Number of TF genes	Number of TF genes with unique domains	% of TF genes with unique domain
2	335	87	25.97%
3	182	38	20.88%
4	135	24	17.78%
5	97	12	12.37%
6	60	3	5%
7	38	5	13.16%
8	24	0	0
9	19	3	15.79%
10	12	0	0
11	9	0	0
12	10	2	20%
13	4	0	0
14	5	1	20%
15	4	0	0
16	3	1	33.33%
17	4	0	0
18	1	0	0
19	1	0	0
22	1	0	0
36	1	0	0

4.2.3 Domains with DNA-Binding Function

Further analyses of transcript and domain relationships are restricted to mouse TF genes with 2 transcripts only. Such genes are in total 335. A total of 87 of these genes have transcripts with unique domains. Further analyses focus on these 87 genes. The transcripts coded by these 87 genes bring in total 147 unique domains.Table 4.9 shows further functional analyses on these domains.

These genes are analyzed in terms of the functions of their unique domains. In particular, DNA binding properties of these domains are investigates. The total number of domains reported in SMART and Pfam for TF genes which produce 2 transcripts with unique domains and the number of unique domains with DNA binding ability is shown in Table 4.9. The percentage of such TF genes is also shown.

Table 4.9: Number and percentage of domains with DNA binding ability from mouse TF genes with 2 transcripts.

Database	Total number of unique domains	Number of domains with DNA binding ability ability	Percentage
Pfam	89	43	48.31%
SMART	58	33	56.90%

As evident from Table 4.9, substantial portions of unique domains analyzed within Pfam and SMART have DNA binding ability.

Lastly, the genes coding for these unique domains are analyzed. Number of TF genes with 2 transcripts, which have DNA-binding ability accounts for about 52% of the genes analyzed in this category, as shown in Table 4.10. This number is important since it shows that more than 50 percent of proteins, which are products of TF genes with 2 transcripts, have DNA binding ability. This result indicates that at least half of the unique domains introduced by different transcripts of a mouse TF gene deliver DNA binding ability.

Table 4.10: Number and percentage of mouse TF genes with 2 transcripts which have DNA-binding ability.

Total number of genes with unique domains	Number of TF genes with unique domains that show DNA binding	Percentage
87	45	51.72%

4.3 Rat TF Transcript Analysis

In the rat genome, a total of 1150 TF genes are analyzed. Figure 4.3 shows the distribution of total number of transcripts per gene. In this figure, x-axis represents the total number of transcripts sequenced per gene. The y-axis represents the total number of TF genes. As evident from this figure, the majority of rat TF genes have 1 or 2 transcripts. As the number of transcripts sequenced per gene increase, the total number of genes decrease.

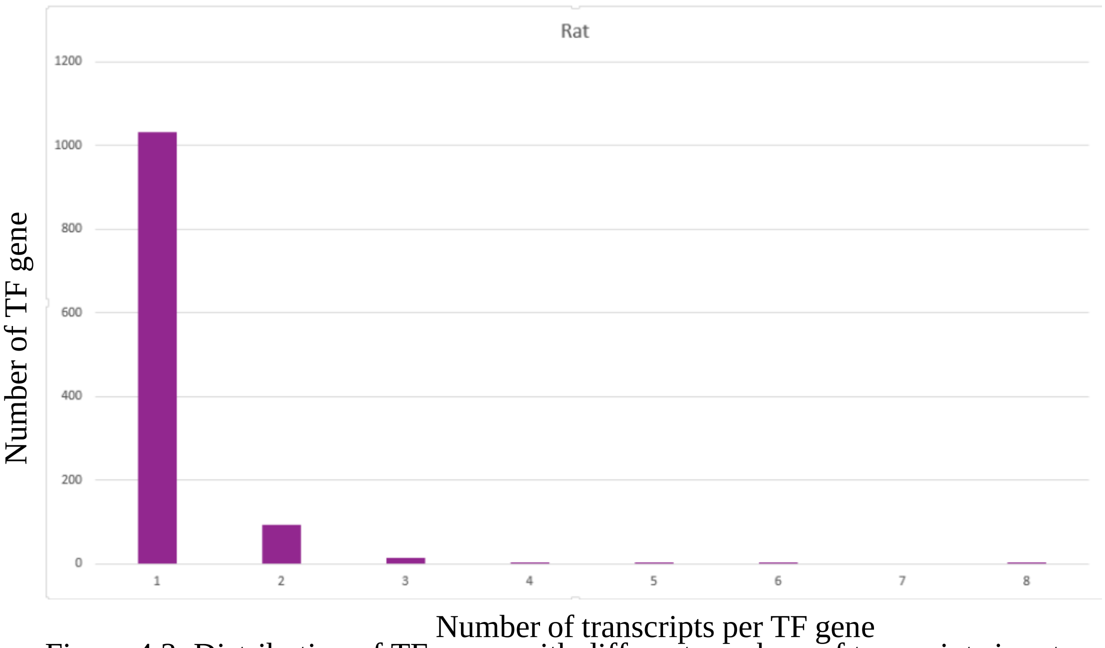


Figure 4.3: Distribution of TF genes with different numbers of transcripts in rat.

4.3.1 TF Gene Categories

Table 4.11 shows the number of TF genes in rat that produce one or more transcripts. Unlike human and mouse, the majority of rat TF genes have only 1 transcript.

Table 4.11: Distribution of TF genes for rat.

Number of TF genes with	Number	Percentage
One transcript only	1033	89.83%
Multiple (two or more) transcripts	117	10.17%

Of these genes with multiple transcripts, the genes that have 2 transcripts i.e. a total of 94 genes are analyzed in terms of their protein coding and their protein structures. This is further explained in the following sections.

4.3.2 Number of TF genes with unique domains

Table 4.12 shows the number of TF genes with multiple different number of transcripts for rat. In addition, this table provides an analysis for the protein domains coded by these transcripts. For each gene, the total number of transcripts available are compared with one another, and unique domains that are present in only one transcript and not the other are reported. As shown in Table 4.12, many genes have transcripts with unique domains. For example, this is the case for 20% of genes with 2 transcripts only and 7% of genes with 3 transcripts only. It would be expected that the percent genes with unique domains would increase as the transcript numbers increase. However, this is not the case. The reason for this is because most of the genes with higher number of transcripts actually have short sequences such as ESTs, and not full-length transcripts.

Table 4.12: Number of rat TF genes with unique domains that is present in only one transcript.

Number of transcripts per TF gene	Number of TF genes	Number of TF genes with unique domains	% of TF genes with unique domain
2	94	19	20.21%
3	14	1	7.14%
4	4	0	0
5	3	0	0
6	1	1	100%
8	1	0	0

4.3.3 Domains with DNA-Binding Function

Further analyses of transcript and domain relationships are restricted to rat TF genes with 2 transcripts only. Such genes are in total 94. And 19 of these genes have transcripts with unique domains. Further analyses focus on these 94 genes. The transcripts coded by these 94 genes bring in total 45 unique domains.Table 4.13 shows further functional analyses on these domains.

These genes are analyzed in terms of the functions of their unique domains. In particular, DNA binding properties of these domains are investigates.The total number of domains reported in SMART and Pfam for TF genes which produce 2 transcripts with unique domains and the number of unique domains with DNA binding ability is shown in Table 4.13. The percentage of such TF genes is also shown.

Table 4.13: Number and percentage of domains with DNA binding ability from rat TF genes with 2 transcripts.

Database	Total number of unique domain	Number of DNA binding ability	Percentage
Pfam	27	11	40.74%
SMART	18	8	44.44%

Unlike human and mouse data, more than half of the domains analyzed in rat do not reveal DNA binding property. The reason for this relies in the fact that a lesser number of rat genes and a lesser numbers of rat transcripts are available for the study. It is expected that this number would rise as more transcripts are sequenced from rat TF genes.

Lastly, the genes coding for these unique domains are analyzed. Number of TF genes with 2 transcripts, which have DNA-binding ability accounts for about 37% of the genes analyzed in this category, as shown in Table 4.14.

Table 4.14: Number and percentage of rat TF genes with 2 transcripts which have DNA-binding ability.

Total number of genes with unique domains	Number of TF genes with unique domains that show DNA binding ability	Percentage
19	7	36.84%

Statistical analysis as described in section 3.3.4 showed that the percentage of TF genes with unique domains that show DNA binding did not differ by species, χ^2 (2, N = 212) = 1.457, p = 0.483, χ^2(1, N = 90) = 0.89, p = 0.35.

Chapter 5

CONCLUSION

5.1 Main Findings

Throughout the studies described in this thesis, several main findings stand out:

i. In all three genomes transcript diversity is documented for TF genes, by comparing the genes with 2 transcripts, and identify their differences in proteins that they code for.

ii. In all three genomes transcripts sequenced from the same gene account for presence of unique domains in the proteins they code for. This could be more confidently stated for human and mouse transcripts as there is more data to be analyzed for these two species.

iii. In human and mouse, the majority of unique domains are responsible for DNA-binding activity.

Overall, in all three genomes, it could be stated that transcript diversity of TF genes result in protein diversity in terms of their domain structures and functional diversity in terms of the DNA-binding ability.

5.2 Future Directions

Further direction of the work described here include:

a) analysis of the specific exon-domain relationships,

b) identification of the source of transcript diversity,

c) expanding this study to other groups of genes and proteins,

d) expanding this study with other genomes.

REFERENCES

[1] Messina D.N., Glasscock J., Gish W., Lovett M., “An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression”, *Genome Res*, Vol 14, 2004.

[2] Van Nimwegen E., "Scaling Laws in The Functional Content of enomes", *Trends Genet.*, Vol 19 (9), PMID 12957540, 2003.

[3] Babu M.M., Luscombe N.M., Aravind L., Gerstein M., Teichmann S.A., "Structure and Evolution of Transcriptional Regulatory Networks," *Curr. Opin. Struct. Biol.*, Vol 14 (3), pp 283 – 291, PMID 15193307, 2004.

[4] Lee T.I., Young R.A., “Transcription of Eukaryotic Protein-Coding Genes,”*Annual Review of Genetics*, Vol. 34, No. 1., pp. 77-137, 2000.

[5] Taneri B., Snyder B., Novoradovsky A., Gaasterland T., “Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific”, *Genome Biology*, Vol 5(10), 2004.

[6] Claverie J.M., Notredame C, *Bioinformatics for Dummies*, Wiley, 2003.

[7] Hogeweg P., “The Roots of Bioinformatics in Theoretical Biology,” *PLoS Comput Biol*, PMID: 21483479, 2011.

[8] Bioinformatics: Introduction and Methods, availeble on https://www.coursera.org /course/pkubioinfo/, retrieved on December 2013.

[9] Jacobs G.H., “Bioinformatics — Computing with Biotechnology and Molecular Biology data,” *BioScience*, Vol. 12(5), pp: 15-18, 2003.

[10] Ouzounis C.A., Valencia A., “Early bioinformatics: the birth of a discipline-a personal view,” *Bioinformatics.* Vol 19(17), pp: 2176-2190 PMID: 14630646, 2003.

[11] Hesper B., Hogeweg P., “Bioinformatica: een werkconcept. Kameleon”, *Leidse Biologen Club(In Dutch.),* Vol. 1(6), pp: 28-29, 1970.

[12] Hogeweg P., and Hesper B., “Interactive instruction on population interactions,” *Comput Biol Med*, Vol. 8, pp: 319-327, 1978.

[13] Watson J.D., and Crick F.H.C., “A Structure for Deoxyribose Nucleic Acid,” *Nature,* Vol. 171 (4356), pp: 737-738, PMID 13054692, 2007.

[14] Elson D., Chargaff E., “On the deoxyribonucleic acid content of sea urchin gametes,” *Experientia,* Vol. 8(4), pp: 143-145, PMID 14945441, 1952.

[15] Genetics home refrences, availeble on http://ghr.nlm.nih.gov/handbook /basics/dna/, retrieved on December 2013.

[16] Dr. Michael Blader, available on http://www.mikeblaber.org, retrieved on December 2013.

[17] DNA structure, available on http://academic.brooklyn.cuny.edu/biology /bio4fv/page/molecular%20biology/dna-structure.html, retrieved on December 2013.

[18] Lewin B., *GENES VIII*, Pearson Prentice Hall,2004.

[19] Mattick J.S., “Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms,” BioEssays, Vol. 25, pp:930-939, 2003.

[20] Control of Gene Expression in Eukaryotes, available on http://sholtoainslie-wordpress.com /portfolio/charts/geneexpress, retrieved on December 2013.

[21] The cell is a messy place: understanding alternative splicing with RNA sequencing, available on http://www.genomesunzipped.org/2010/12/the-cell-is-a-messy-place-understanding-alternative-splicing-with-rna-sequencing.php, retrieved on December 2013.

[22] Eukaryotic mRNA Processing, available on http://oregonstate.edu/dept/ biochem/hhmi/hhmiclasses/bb451/lectnotesgdp/EukaryoticmRNA.html, retrieved on December 2013.

[23] Simple Gene Expression, available on http://highered.mcgraw-hill.com/sites /0072995246/student_view0/chapter3/simple_gene_expression.html, retrieved on December 2013.

[24] Alberts B., Johnson A., Lewis J., *Molecular Biology of the Cell*. New York: Garland Science; 2002.

[25] Transcription complex and enhancers, available on http://highered.mcgraw-hill.com, retrieved on December 2013.

[26] Kolovos P., Knoch T.A., Grosveld F.G., Cook P.R., and Papantonis A., "Enhancers and silencers: an integrated and simple model for their function", *Epigenetics & Chromatin, Vol.5, 2012.*

[27] mRNA Synthesis (Transcription), available on http://highered.mcgraw-hill.com /sites/0072507470/student_view0/chapter3/, retrieved on December 2013.

[28] Rosonina E., Kaneko S., and Manley J.L., "Terminating the transcript: breaking up is hard to do" , *Genes & Dev,* Vol.20, pp: 1050-1056, 2006.

[29] Greive S.J., Peter H., Hippel V, "Thinking quantitatively about transcriptional regulation", *Nature Reviews Molecular Cell Biology,* Vol.6, pp:221-232, 2005.

[30] Translation, available on http://highered.mcgraw-hill.com /sites/0072507470/student_view0/chapter3/, retrieved on December 2013.

[31] Translation and Transcription, available on http://www.rci.rutgers.edu/~uzwiak /GBSummer13/GB101Summer13Lect/GB101Summer_Lect17.htm, retrieved on December 2013.

[32] Cooper G.M., *The Cell: A Molecular Approach,* 2nd edition. Sunderland, 2000.

[33] Protein Production: A Simple Summary of Transcription and Translation, available on http://hubpages.com/hub/protein-production-a-step-by-step-illustrated-guide, retrieved on December 2013.

[34] Raven P., Johonson G., Losos J., Mason K., and Singer S., *Biology,* 8th edition Missouri Botanical Gardens & Washington University, 2008.

[35] RNA splicing: introns exons and spliceosome, available on http://www.nature.com/scitable/topicpage/rna-splicing-introns-exons-and-spliceosome-12375, retrieved on December 2013.

[36] Transcription & Translation: RNA Splicing, available on http://www.dnalc.org /resources/3d/rna-splicing.html, retrieved on December 2013.

[37] mRNA Splicing, available on http://www.hartnell.edu/tutorials /biology/splicing.html, retrieved on December 2013.

[38] Black D.L., “Mechanisms of alternative pre-messenger RNA splicing”. *Annual Reviews of Biochemistry* Vol.72, PP:291–336, PMID 12626338, 2003.

[39] Alternative splicing, available on http://www.eurasnet.info/education /alternate-splicing/what-is-alternate-splicing, retrieved on December 2013.

[40] Claudia G., Cristina V., and Giuseppe B., “Alternative Splicing and Tumor Progression”, *Curr Genomics journal*, Vol.9, pp: 556-570, PMCID: PMC2694562, 2008.

[41] Alternative splicing, available on http://www.genome.gov/Images/EdKit /bio2j_large.gif, retrieved on December 2013.

[42] Types of alternative splicing, available on http://www.tau.ac.il/~gilast /research_as.html, retrieved on December 2013.

[43] Alternative Splicing, available on http://en.wikipedia.org/wiki/File: AlternativeSplicing.png, retrieved on December 2013.

[44] A short amino acid sequence, available on http://medical-dictionary. thefreedictionary.com/monomer, retrieved on December 2013.

[45] Proteins, available on http://cnx.org/content/col11496/1.6 , retrieved on December 2013.

[46] Types of Amino Acids, available on http://naribiochemwiz007.files.wordpress .com/2013/04/amino_acids.gif, retrieved on December 2013.

[47] Perler F.B., Xu M.Q., Paulus H., “Protein Splicing and Autoproteolysis Mechanisms”. *Curr Opin Chem Biol,* Vol.1, PP: 292–299, PMID 9667864, 1997.

[48] Protein primary structure, available on http://www.genome.gov/Pages /Hyperion//DIR/VIP/Glossary/Illustration/amino_acid.shtm, retrieved on December 2013.

[49] Arunan E., Desiraju G.R., Klein R.A., Sadlej J., Scheiner S., Alkorta I., Clary D.C., Crabtree R.H., Dannenberg J.J., Hobza P., Kjaergaard H.G., Legon A.C., Mennucci B., “Definition of the hydrogen bond”, *Pure Appl. Chem.*, Vol. 83, pp. 1637–1641, 2011.

[50] Protein Secondary Structure: α-Helices and β-Sheets, available on http://www. proteinstructures.com/Structure/Structure/secondary-sructure.html, retrieved on December 2013.

[51] Chemistry and the Building Blocks of Life, available on http://www. uic.edu/classes/bios/bios100/lectures/chemistry.htm, retrieved on December 2013.

[52] 3-Dimensional Protein Structures, available on http://www.ncbi.nlm.nih.gov /Class/MLACourse/Original8Hour/Genetics/structure.html, retrieved on December 2013.

[53] Protein Domains, available on http://www.ebi.ac.uk/training /online/course/introduction-protein-classification-ebi/protein-classification /what-are-protein-domains, retrieved on December 2013.

[54] Gu J., and Bourne P.E., *Structural Bioinformatics Second Edition*. Wiley 2009.

[55] Barrie E.S., Smith R.M., Sanford J.C., and Sadee W., "mRNA Transcript Diversity Creates New Opportunities for Pharmacological Intervention", *Molecular Pharmacology,* Vol.81, pp: 620–630, PMCID: PMC3336806, 2012.

[56] LEE T.K.B., "Bioinformatics Analysis Of Alternative Splice", A Phd thesis, National University of Singapore, 2005.

[57] Casado Vela J., Lacal J.C., Elortza F., "Protein chimerism: novel source of protein diversity in humans adds complexity to bottom-up proteomics. Proteomics", Vol. 13(1), pp: 5-11, 2013.

[58] The National Center for Biotechnology Information NCBI, available on http://www.ncbi.nlm.nih.gov/, retrieved on December 2013.

[59] Ensembl Genome Browser, available on http://www.ensembl.org/index.html, retrieved on December 2013.

[60] BioMart, available on http://www.ensembl.org/biomart/martview /bb9b745659f6b988a0420d821523a4fb, retrieved on December 2013.

[61] Simple Modular Architecture Research Tool, available on http://smart.embl-heidelberg.de/, retrieved on December 2013.

[62] Pfam, available on http://pfam.sanger.ac.uk/family/PF03474, retrieved on December 2013.

[63] Eric W., "Database Resources of The National Center for Biotechnology Information ", *Nucleic Acids Research*, 2010, PMCID: PMC2808881.

[64] Jo M.E., Jim O., *The NCBI Handbook,* National Center for Biotechnology Information (US), 2002.

[65] Hubbard T., Barker D., Birney E., "The Ensembl genome database project", *Nucl. Acids Res,* Vol.30, pp:38-41, 2002.

[66] Ensembl tools, available on http://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/how-export-sequence-and-download-data/ensembl-tools, retrieved on December 2013.

[67] BioMart Taturial, available on http://www.ensembl.org/info/data/biomart.html, retrieved on December 2013.

[68] Schultz J., Milpetz F., Bork P., and Ponting C., "SMART, a simple modular architecture research tool: Identification of signaling domains", Proc Natl Acad Sci USA, Vol.95, pp: 5857-5864, PMCID: PMC34487, 1998.

[69] An introduction to the Pfam protein families database, available on http://pid.nci.nih.gov/2011/110913/full/pid.2011.3.shtml, retrieved on December 2013.

[70] Astrahan M., Blasgen M., Chamberlin D., Eswaran K., Gray J., Griffiths P., King W., Lorie R., McJones P., Mehl J., Putzolu G., Traiger I., Wade B., Watson V., " System R: a relational approach to database management", ACM Transactions on Database Systems, Vol.1, pp: 97–137, 1976.

[71] Codd E.F., "A Relational Model of Data for Large Shared Data Banks", IBM Research Laboratory, Vol.13, 1970.

[72] Date C.J., *An introduction to database systems,* 6th edition, Addison-Wesley, 1995.

[73] Conceptual Modeling using the Entity-Relationship Model, available on http://www.cs.ucdavis.edu/~green/courses/ecs165a-w11/2-er.pdf, retrieved on December 2013.

[74] Peter P.S.C., "The Entity-Relationship Model-Toward a Unified View of Data", *ACM Transactions on Database Systems,* Vol. 1, pp: 9-36, 1976.

[75] Codd E.F., "Further Normalization of the Data Base Relational Model", IBM Research Report, 1971.