GENOME DESIGN IN EUKARYOTES

NANYANG TECHNOLOGICAL UNIVERSITY

LI PENG (G0602350B)

SCHOOL OF MECHANICAL AND AEROSPACE ENGINEERING

A thesis submitted to the Nanyang Technological University in fulfillment of the requirement for the degree of Master of Engineering

2010

ABSTRACT

The availability of complete genome sequences for many eukaryotic organisms continues to contribute towards a better understanding of their genome design and evolution. This investigation involves computational analysis of genome architecture of 6 eukaryotic genomes (4 vertebrate: *H.sapiens, P.troglodytes, M.musculus, D.rerio*; 2 invertebrate: *C.elegans, D.melanogaster*). We further analyzed the drug targets, that is, proteins in the human genome with FDA (Food and Drug Administration) approved drugs using various parameters such as protein interacting partners, number of exons, number of pathways, number of tissues and protein family to find if there is any co-relation between these parameters and the targetability of the protein. It was observed that proteins from single exonic genes are more likely to have an FDA approved drug.

These data have implications in understanding eukaryotic genome design and may also contribute in drug target selection which is the most important step in drug discovery.

Further, a database was constructed on discordant introns. These investigations will help us in understanding eukaryotic genome design.

ACKNOWLEDGEMENTS

First, I would thank my supervisor Prof. Zhong Zhaowei and my ex-supervisor Prof. Meena Sakharkar, for their continuous support and encouragement throughout the conceptualization and carrying out of this work. It would not have been possible without their patience and support. They showed me different ways to approach a research problem and the need to be persistent to accomplish the goal. Without their help and direction, I may not have been able to finish this project.

I would also like to thank the Head of Mechatronics and Design, Prof. Gerald Seet and the School of Mechanical and Aerospace Engineering for providing the necessary infrastructure to carry out this work. The efforts of Mr. Justin, Mr. Teo Hai Beng, Mr. Chia, Mr. Koh and Mr. Soh of ADAMs laboratory were noteworthy in completion of this thesis.

Last, but not least, I would like to thank my family: my parents, parents in law, my wife for their unconditional support and encouragement for pursuing this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 FUNDAMENTAL OF COMPUTATIONAL RESEARCH ON GENE

1.1 Gene

The smallest functional unit of inherited information is a gene (Morgan, 1917). A gene is a DNA sequence and most genes contain information for making specific proteins [Figure1.1]. Each DNA molecule is composed of two polynucleotide strands twisted around each other to form a double helix (Watson and Crick, 1953) [Figure 1.2]. The polynucleotide is made of four types of nucleic acid bases (Adenine, Thymine, Guanine and Cytosine represented as A, T, G and C respectively). Each strand has a chemical polarity, described as going from a 5' end to a 3' end, and this is based on the position of the carbon atom on the pentose ring to which phosphate groups bind in either direction [Figure 1.3]. The genetic code is read as a series of codons, each of which consists of three base pairs (bp), which in turn corresponds to a single amino acid (Crick, 1968)

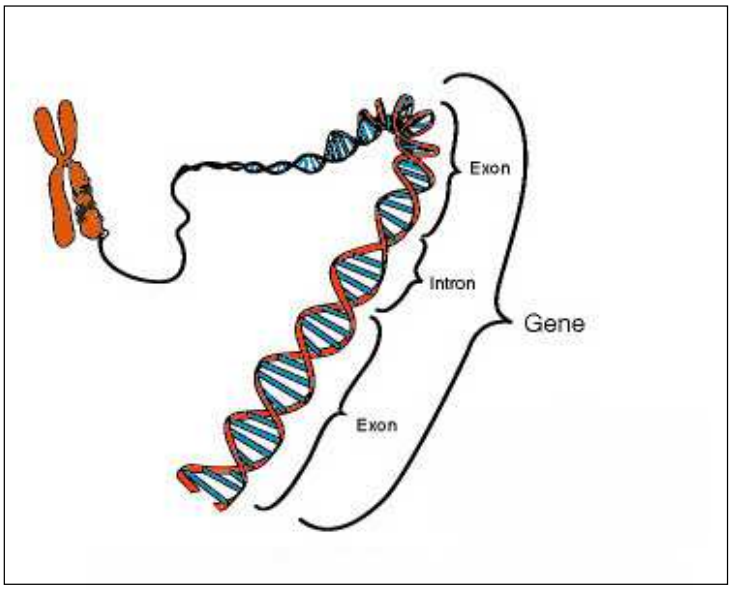

Figure 1.1
Genes are pieces of DNA and most genes contain the information for making a specific protein. The regions that code for the protein are called exons. Long regions of DNA called introns that have no apparent protein coding function separate exons in eukaryotes. This picture is reproduced from the glossary of genetic terms at http://www.nhgri.nih.gov/DIR /VIP/glossary/pub-glossary.cgi.

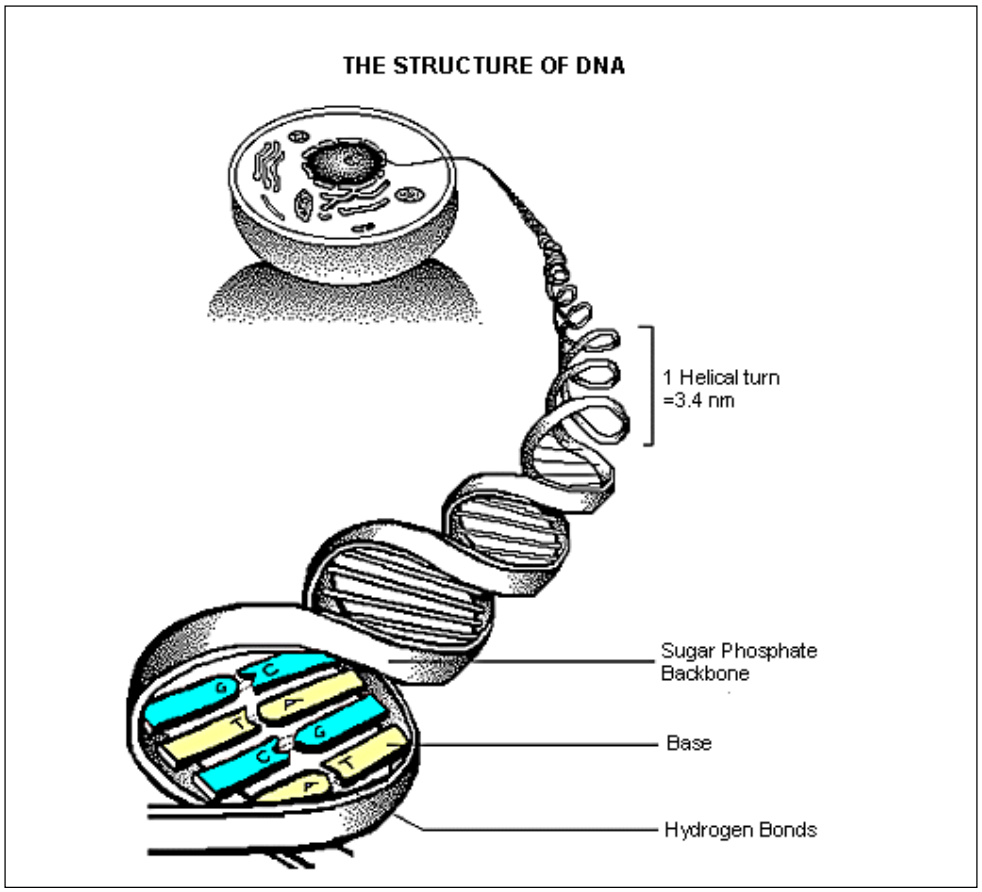

Figure 1.2
The structure of DNA is illustrated by a right-handed double helix, with about 10 nucleotide pairs per helical turn. Each spiral strand, composed of a sugar phosphate backbone and attached bases, is connected to a complementary strand by hydrogen bonding between paired bases, adenine (A) with thymine (T) and guanine (G) with cytosine (C). James Watson and Francis Crick first described this structure in 1953. This picture is reproduced from the glossary of genetic terms at http://www.nhgri.nih.gov/DIR /VIP/glossary/pub-glossary.cgi.

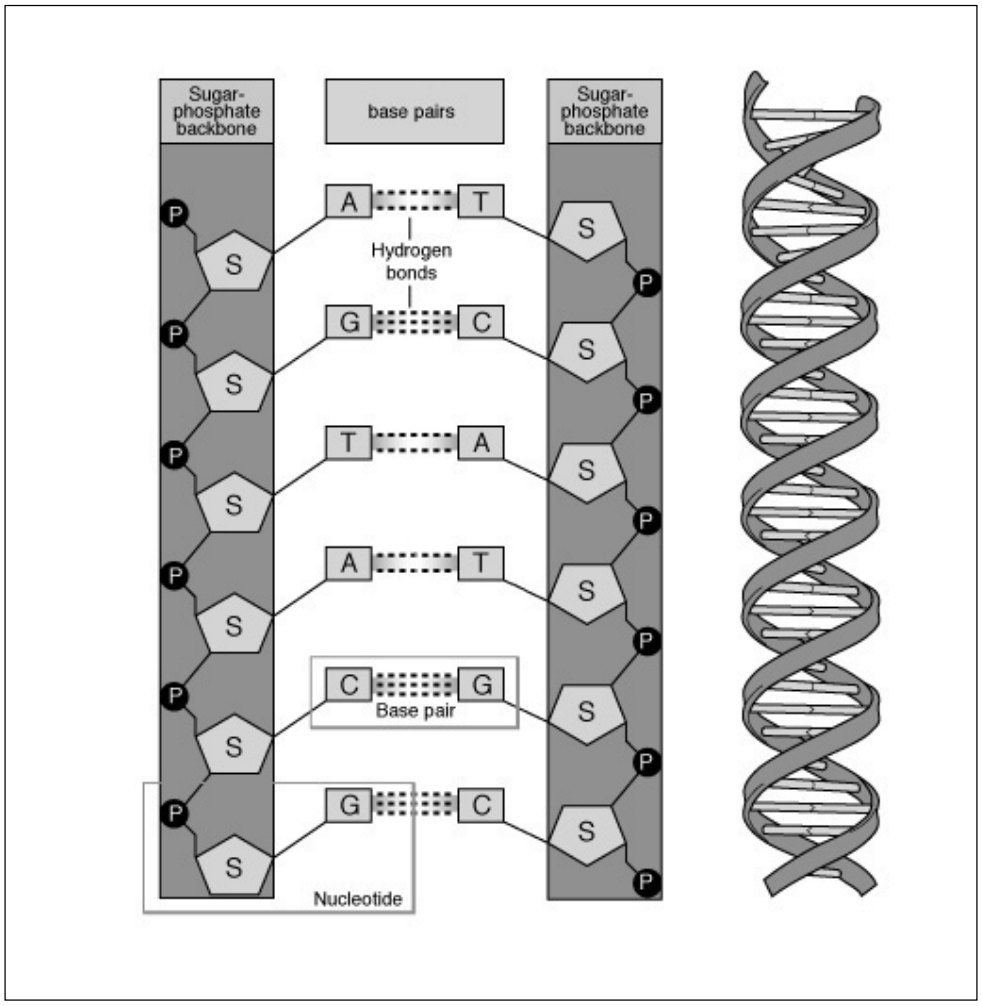

Figure 1.3
The chemical structure of DNA is illustrated in detail. Adenine (A) and thymine (T) are connected by two hydrogen bonds (non-covalent) while guanine (G) and cytosine (C) are connected by three hydrogen bonds. S = sugar; P = phosphate. This picture is reproduced from the glossary of genetic terms at http://www.nhgri.nih.gov/DIR /VIP/glossary/pub-glossary.cgi.

1.2 Eukaryotic Gene

In eukaryotes, the transcribed sequence (synthesized RNA) is further divided into exons (coding sequences) and introns (intervening non-coding sequences). This is illustrated in Figure 1.4a and Figure 1.4b.

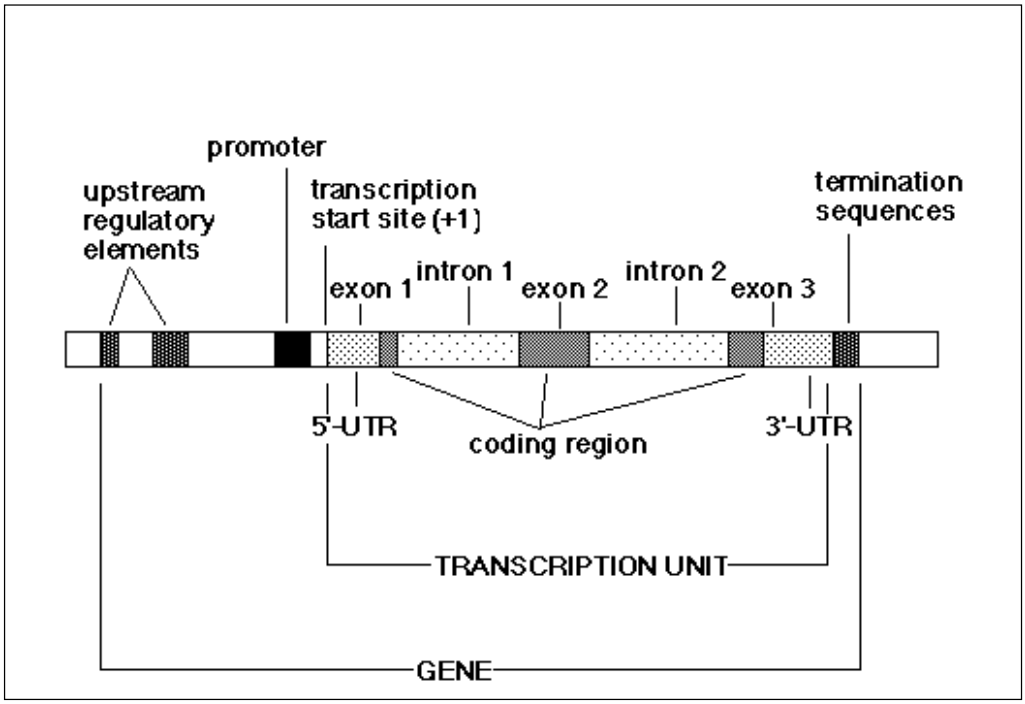

Figure 1.4a
Schematic representation of an eukaryotic gene showing different components required for producing the primary transcript. The primary transcript is much longer than the mRNA coded by the gene because of the regulatory elements required for transcription, processing of the 3' end and the intervening sequences (introns).

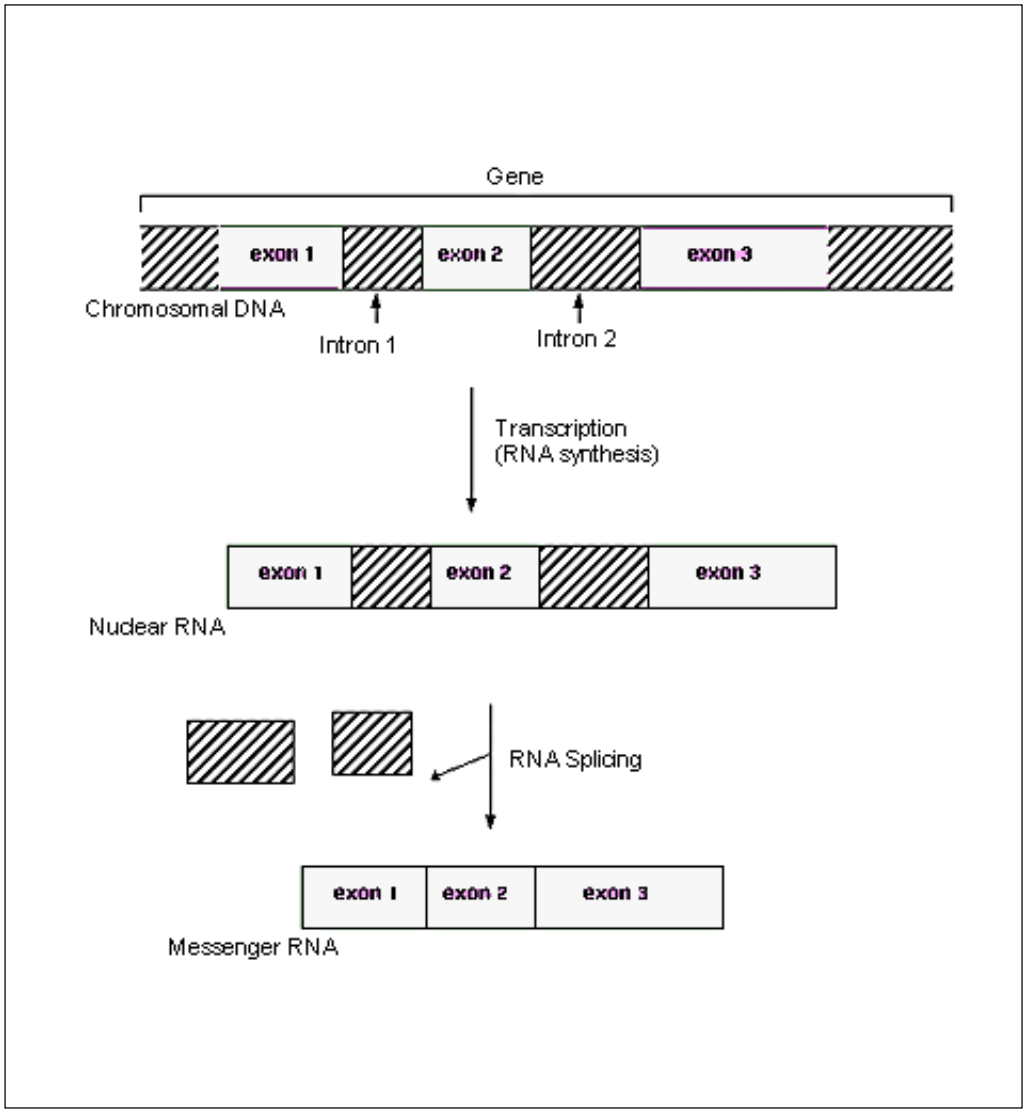

Figure 1.4b
In chromosomes, DNA acts as a template for the synthesis of RNA in a process called transcription. DNA is transcribed to produce pre-processed RNA which after processing is ready to be translated into a protein. In many eukaryotic genes, the DNA sequence coding for proteins, or ‘‘exons’’, may be interrupted by stretches of non-coding DNA, called ‘‘introns’’. This process involves removal of introns and joining of exons. The edited sequence is called ‘‘messenger RNA’’ or mRNA. The mRNA, which carries the gene's instructions, dictates the production of proteins by the ribosomes. This picture is reproduced from the glossary of genetic terms at http://www.nhgri.nih.gov/DIR /VIP/glossary/pub-glossary.cgi.

1.3 Exons

The genetic unit of function in the gene that corresponds to a polypeptide chain is called the exon. However, in genes that do not code for a functional protein product such as the tRNA genes of yeast and the rRNA genes in Drosophila, and the viral messages from adenovirus, Rous sarcoma virus and murine leukaemia virus, the primary RNA transcript contains internal regions that are excised during maturation. The final tRNA or messenger is a spliced product. When interrupted genes code for rRNA or tRNA the exons do not have a protein coding function (Gilbert, 1978).

1.4 Introns

The term intron (intragenic regions) as proposed by Gilbert in 1978 represents the genetic unit that will be lost in the mature messenger RNA (Gilbert, 1978). The unexpected extra DNA in eukaryotic cells, the excess of DNA over that needed to code for products defined genetically, is ascribed as introns. However, some eukaryotic protein coding genes are not interrupted by introns. For example, the human G-protein coupled receptors are predominantly intronless (Gentles and Karlin, 1999). Some other intronless genes include the multiple genes encoding the histone proteins (Schaffner *et al.*, 1978) and the uninterrupted genes coding for interferon proteins in vertebrates during viral infections.

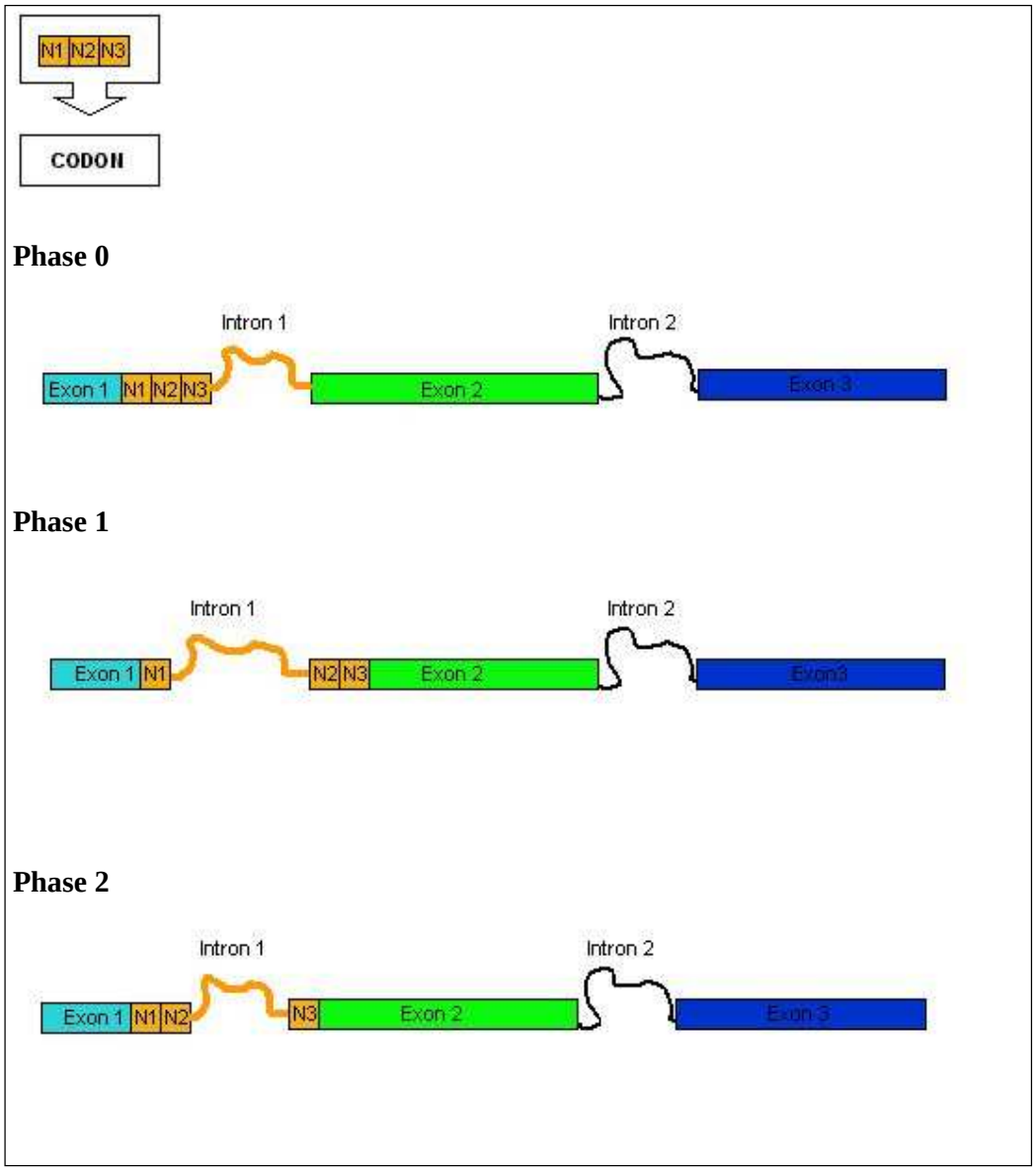

Figure 1.5
Types of introns based on their position relative to the reading frame. Phase 0, Phase 1 and Phase 2 introns are shown schematically. N1, N2 and N3 are nucleotides representing a codon.

Properties of Introns

(1) Introns are intervening non-coding sequences in eukaryotic genes (Gilbert, 1978).

(2) Introns are found in nuclear genes coding for proteins, rRNA, tRNA, mitochondrial genes of lower eukaryotes, chloroplast genes, T4 phage, archaebacteria and rarely in eubacterial genes (Gilbert, 1978).

(3) Introns invariably begin and end with highly preserved splice site sequences [beginning and ending with GT and AG respectively (Dibb, 1991) or occasionally begin with AT and end with AC] (Nilsen, 1996; Tarn and Steitz, 1996).

(4) Intron positions correlate with module boundaries in ancient proteins (de Souza *et al.*, 1996).

(5) Introns are involved in many cellular functions including recombination and gene regulation (Carvalho and Clark, 1999).

(6) The evolutionary advantage of having introns is the potential to specify more than one polypeptide from a single gene. It turns out that for a substantial minority of genes, the primary transcript can be spliced together in more than one way, a process known as alternative splicing (Hanke *et al.*, 1999).

(7) Introns participate in mutagenesis and gene reorganization leading to diseases (Janssen *et al.*, 2000).

(8) Generally, as one proceeds down the evolutionary ladder fewer and fewer protein-coding genes are interrupted and the introns that do exist tend to be shorter. For instance, the human and the rat cytochrome c genes are interrupted by introns while the yeast cytochrome c gene lacks introns. The Drosophila alcohol dehydrogenase gene has introns while the comparable

yeast gene does not. Less than 4% of yeast genes contain introns and introns account for less than 1% of the entire genome (Patthy, 1999).

(9) Genome size increase with intron size (Deutsch and Long, 1999).

(10) Occasionally, introns assume a function within the cell as templates not for regular RNA but for RNA that has another role and the exons in such a gene are silent (Tycowski *et al.,* 1996).

(11) A number of eukaryotes require at least one intron for their transcripts to accumulate to maximal levels (Hamer and Leder, 1979; Buchman and Berg, 1988). To account for this intron dependent RNA accumulation, it has been suggested that introns contribute to 3' end processing efficiency (Huang and Gorman, 1990; Pandey *et al.,* 1990; Nesic and Maquat, 1994), nuclear stability and nucleocytoplasmic transport (Ryu and Mertz, 1989) or transcriptional enhancement (Chung and Perry, 1989).

1.5 Evolution of Exons and Introns

Shortly after the discovery of split genes, it was realized that the existence of introns may have a necessary role in the biology of the eukaryotic cell and might have dramatic consequences on protein evolution (Gilbert, 1978). It was pointed out that recombination within introns could assort exons independently, and middle repetitious sequences in introns may create hotspots for recombination to shuffle the exonic sequences. Two types of hypotheses explain the presence of introns in most eukaryotic protein-coding genes and their absence from prokaryotes.

1.5.1 “Intron Early” Hypotheses

To suggest that introns are ‘‘early’’ is to suggest that protein-coding genes of the cenancestor (the most recent common ancestor of archaebacteria, eubacteria and eukaryotes) had a genome full of introns, and that ‘‘streamlining’’ (systematic loss of introns) accounts for the existence of intronless and intron-poor genomes. The ‘‘introns early’’ hypotheses assumed that introns and RNA splicing are the relics of the RNA world and the ‘‘genes in pieces’’ organization of the eukaryotic genome is the original, ancestral form (Darnell, 1978; Doolittle, 1978; Darnell and Doolittle, 1986; Gilbert, 1986). According to this view, eukaryotes retained introns and the genetic plasticity of the primitive ancestors of all cells. On the other hand, bacteria gained increased efficiency by eliminating their introns. Supporters of the introns-early hypotheses assume that the introns of all protein-coding genes reflect the assembly of these genes from pieces; that exons do indeed correspond to building blocks from which all the genes were assembled by intronic recombination (Gilbert and Glynias, 1993).

One version of ‘‘introns-early’’, the ‘‘exon theory of genes’’ (Gilbert, 1987), proposes that exons date back not only to the cenancestor, but much further, to the very first protein-coding genes - primordial mini genes that were later assembled into modern-sized genes by mixing and matching (‘‘exon shuffling’’) (Blake, 1978; Blake, 1979; Blake, 1983; Gilbert, 1978). The “exon theory of genes” pursues the idea that the shuffling of a limited number of exons whose products are modular folding elements is an extremely potent way to generate a large diversity of protein structures (de Souza *et al.*, 1996). It suggests that the ancient proteins were essentially aggregates of short

polypeptides perhaps 15-20 amino acid residues in length and that the genes were initially assembled from small exons of this size.

1.5.2 "Intron Late" Hypotheses

In contrast with ''intron early theory'', the ''introns late'' theories suggest that the prokaryotic genes resemble the ancestral ones and that the introns were inserted later in genes of eukaryotes (Crick, 1979; Cavalier, 1985; Cech, 1985; Orgel and Crick, 1980; Sharp, 1985). The exon-intron structure of eukaryotic protein-coding genes is evolving: introns are continually inserted into (as well as removed from) genes. The actual mechanisms of insertion, propagation of some self-splicing introns have been analyzed in detail and the mechanisms responsible for the insertion of spliceosomal introns are also becoming clear (Dujon, 1989; Lambowitz and Belfort, 1989; Perlman and Butow, 1989; Morl and Schmelzer, 1990; Belfort, 1991; Belfort, 1993; Lambowitz, 1993; Mueller *et al.*, 1993; Grivell, 1994; Patthy, 1995).

Since introns themselves are subject to evolution, it is clear that exon shuffling has been evolving in parallel with the evolution of introns. Previous argument shows that the introns suitable for exon shuffling appeared at a relatively late stage of evolution, therefore, exon shuffling could not play a major role in the construction of ancient proteins (Patthy, 1987; Patthy, 1991). The self-splicing introns of the RNA world that could be present at the time the first proteins were formed are practically unsuitable for exon shuffling by intronic recombination: such self-splicing introns already encode the essential splicing function, therefore their sequence is not tolerant to intronic recombination (Patthy, 1987; Patthy, 1991; Patthy, 1994). Exon-shuffling could become significant only with the appearance of spliceosomal introns. These introns

play a negligible role in their own excision; therefore, intronic recombination is less likely to produce recombinant introns that are deficient in splicing. Furthermore, the nonessential parts of spliceosomal introns could accommodate large segments of middle repetitious sequences, further increasing the chances of intronic recombination. Since spliceosomal introns evolved relatively recently from group II self-splicing introns (Cech, 1986; Jacquier, 1990; Cavalier, 1991; Copertino and Hallick, 1993; Sharp, 1994) and are restricted in their evolutionary distribution (Cavalier, 1991; Logsdon, 1991; Palmer and Logsdon, 1991) exon-shuffling could play a major role only in the construction of ‘‘younger’’ proteins (Patthy, 1987; Patthy, 1991; Patthy, 1994; Patthy, 1995; Patthy, 1996).

1.5.3 Different Approaches to Test Models for Intron Evolution

Position of exon-exon boundaries in proteins of known structure

Immunoglobulin

If introns are ancient-relics of the way proteins were built as modules, exon-exon boundaries should be found at the boundaries between domains. Immunoglobulin exons correspond to domains and functional elements [Figure 1.7] (Sakano *et al.*, 1979; Early *et al.,* 1979). Introns separate the three constant region domains CH1, CH2 and CH3 in the cases of gamma-1 and the alpha heavy chains. Separate exons encode the hinge region and the sequence of 15 amino acid residues that lies between CH1 and CH2. The former exon provides a flexible point in the molecule and the latter exon includes the cysteines that form the disulphide connections between the two heavy chains.

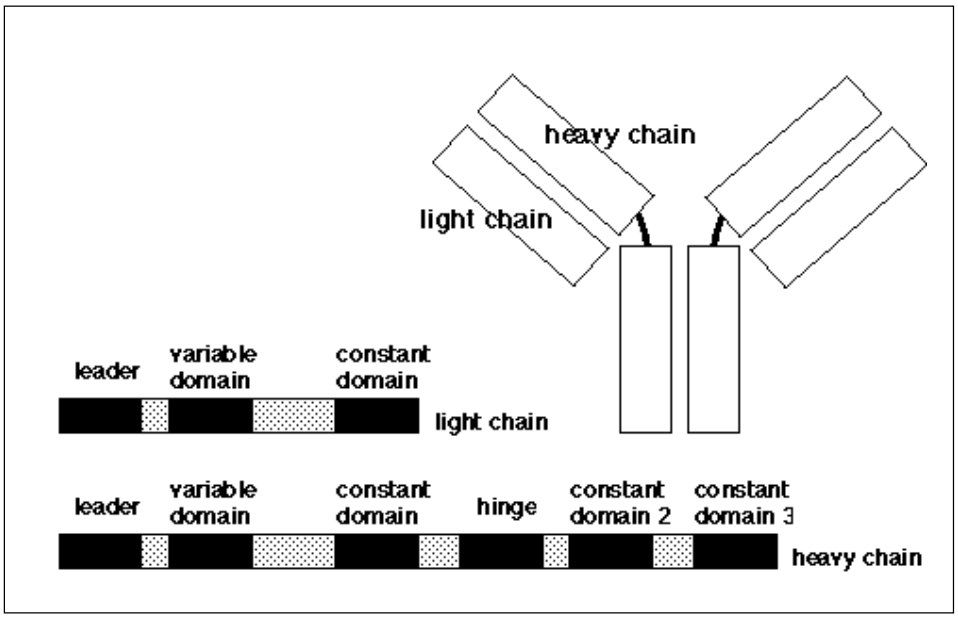

Figure 1.6
Functional elements (variable domains, constant domain, hinge and leader) in the immunoglobulin protein product correspond to exons in the gene. The dark shaded boxes correspond to exons and the dotted boxes correspond to introns.

Globins

Globins have a simpler gene structure. Haemoglobin when colored by exons shows no obvious relationship between structural modules and exon boundaries [Figure 1.8] (Tilghman *et al.*, 1978; Jeffreys and Flavell, 1977; Leder *et al.*, 1978).

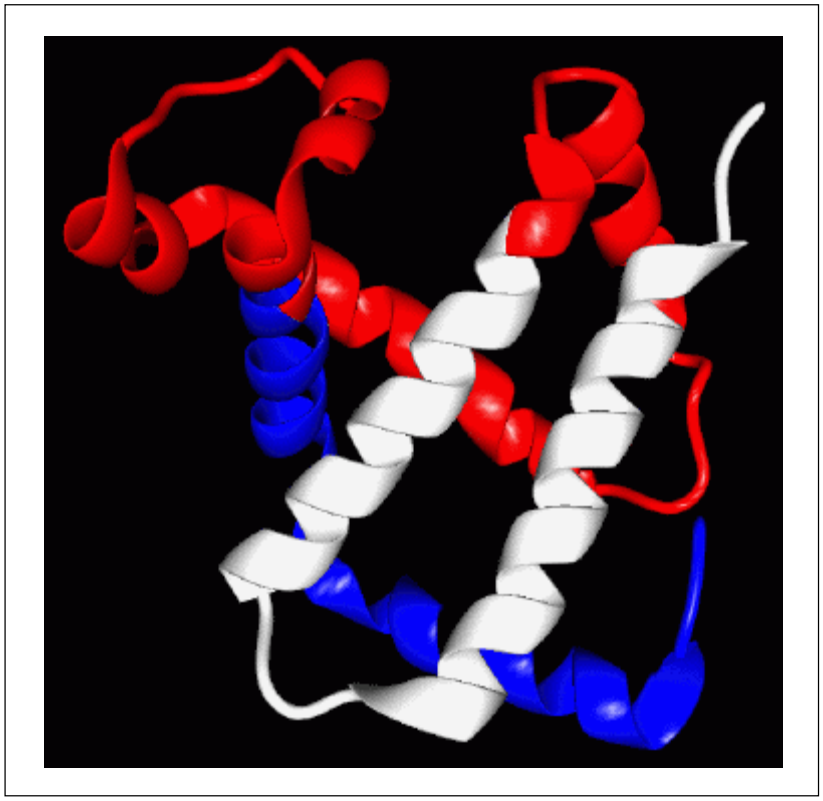

Figure 1.7
Haemoglobin is colored based on exon boundaries (the three exons are colored in blue, red and white). There is no obvious relationship between structural modules and exon boundaries.

Construction of Phylogenetic Trees for Evolutionarily Related Genes and Identifying The Differences in Intron-Exon Structure

If introns are recent, one should see lineages with introns present at different locations. The following picture is a diagrammatic representation of distribution of introns in Actins [Figure 1.9]. Actins are ubiquitous proteins conserved in evolution and an analysis of their gene structures from various organisms has revealed that there may be at least 25 intron positions distributed at different positions in the coding regions. A comparison of intron positions from a wide range of organisms from that of yeast to human actins shows that introns could be ancestral in origin. The conservation in the observed intron patterns within the different tissue types hints at a possible functional significance of introns in present day actin genes (Bagavathi and Malathi, 1996).

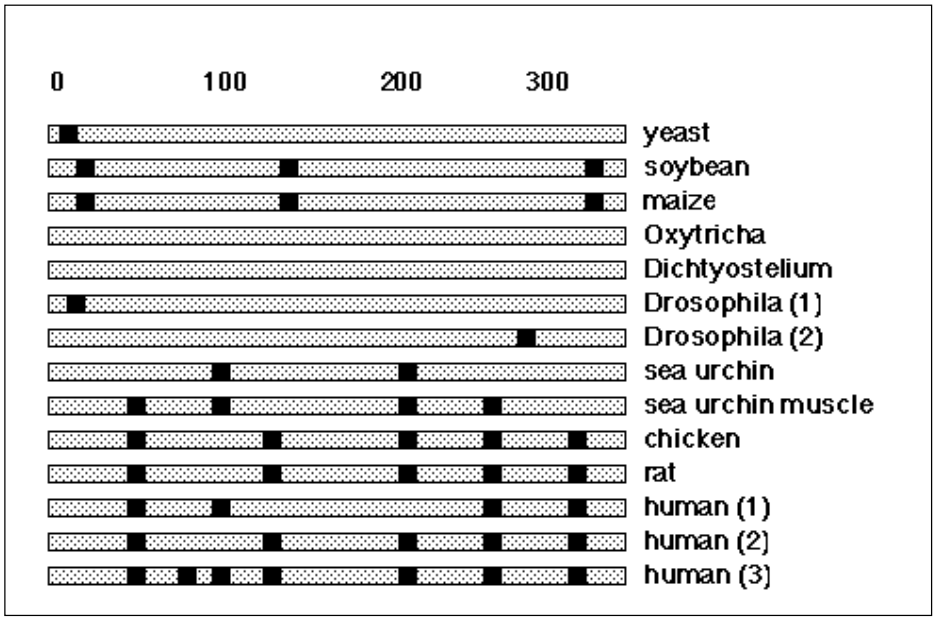

Figure 1.8
The distribution of introns in actin gene family is illustrated schematically. Black bars represent introns. The intron locations are shown by codon positions at the top. This picture is based on Bagavathi S (1996).

1.6 Eukaryotic Gene Structure Prediction

A brief summary of gene finding approaches in eukaryotes has been reported elsewhere (Stormo, 2000).Gene structure identification includes 2 steps:

(1) Type of information used by the program

(2) Algorithm used for prediction to combine that information into a coherent prediction

Type of Information Used by The Program

(1) Signals

(2) Content Measures

(3) Similarity Measures

Signals

Signals involves recognition of spliced sites, i.e, the donor site and the acceptor site. If signals could be detected, the problem in finding the codeing region from the open reading frame could be solved.

Content Measures

Content Measures involoves identification of statistical characteristics that can help distinguish coding region from non-coding region.

Similarity Measures

Similarity measures can be employed to identify coding region and non-coding region compound of genomic DNA.

Algorithm that Is Employed to Combine The Information into A Coherent Prediction

Sequence similarity is assessed typically by sequence comparison algorithms that follow two main approaches.

(1) Dynamic Programming (Bertsekas, 1995; Gelfand and Roytberg, 1993; Snyder and Stormo, 1993)

(2) Hidden Markov Models (Churchill, 1989)

Such algorithms can be used to combine the similarity measure information into a coherent prediction.

1.7 Efficiency of Gene Structure Prediction Programs

Current software tools are moderately effective in predicting genetic structure (exons, introns, intergenic regions and complete genes) from raw DNA sequence data. Improvements in accuracy and speed are needed to deal with the increasing volume of data from large scale sequencing projects. A comparison for different gene structure prediction programs is shown in [Table 1.1]. The performance comparison shown below is for the Burset /Guigo test set (Burset and Guigo, 1996) (570 vertebrate genes, average coding proportion 0.21). n/a = not applicable.

Program	Sn	Sp	Sq	CC	AC	XSn	XSp
GRPL(Hu)	0.93	0.93	0.984	0.91	0.91	0.76	0.79
GRPL (Hu)+	0.97	0.97	0.990	0.96	0.96	0.81	0.85
GENSCAN	0.93	0.93	n/a	0.92	0.91	0.78	0.81
Genie	0.76	0.77	n/a	n/a	0.72	0.55	0.48
GRAIL2	0.72	0.87	n/a	0.76	0.75	0.36	0.43
GeneID	0.63	0.81	n/a	0.65	0.67	0.44	0.46
Xpound	0.61	0.87	n/a	0.69	0.68	0.15	0.18
GeneID+	0.91	0.91	n/a	0.88	0.88	0.73	0.70
GeneParse3	0.86	0.91	n/a	0.85	0.86	0.56	0.58

Table 1.1
A comparison for different gene structure prediction programs is shown (reproduced from *Hooper, 2000*). Sp = specificity, Sn = sensitivity, CC = correlation coefficient, XSn = exon sensitivity, XSp = exon specificity, AC and CC are similar. The detailed description of each parameter is available elsewhere (Hooper, 2000).

1.8 Exon - Intron Distribution in Crown Eukaryotic Genomes

Annotation of genes is an integral part of every genome-sequencing project. However, a large number of uncharacterized DNA sequences are generated by eukaryotic genome projects. When one determines the intron-exon structure of a newly characterized gene, one wonders if it is a known structure or if it represents an entirely novel structure. This involves exact determination of gene structures such as coding and non-coding regions, promoters and transcription regulatory elements. Thus, it becomes essential to develop algorithms for computational gene finding. Developing sensitive computational methods to find genes and open reading frames in eukaryotic genome sequences is an important task in genomics studies. It depends on the complete statistical description and understanding of intron-exon organization in

eukaryotic genes. An updated statistical description of intron-exon structures has been missing and is important for the theoretical study of the origin and evolution of genes and genomes. Recently, the statistical distribution of introns and exons in 10 eukaryotic model organisms was reported (Deutsch and Long, 1999). The study reports the following:

(1) A eukaryotic gene on average contains 3.7 introns per Kb protein coding region.

(2) Most exons are 90-120 nucleotides (30-40 amino acid) residues in the dataset.

(3) Most introns are 40-125 nucleotides long in the dataset.

(4) Genome size seems to be increased with total intron length per gene. For example, invertebrate introns are smaller than those of human genes, whion the other hand yeast introns are shorter than invertebrate introns.

(5) Introns of size smaller than 50 base pair are significantly less frequent than longer introns, possibly resulting from a minimum intron size requirement for intron splicing.

1.9 Genbank – The Pre-Eminent Nucleotide Sequence Database

Overview

GenBank is the National Institute of Health's genetic sequence database and is an annotated collection of all publicly available nucleotide and protein sequences. The unit records in GenBank represent single, contiguous stretches of DNA or RNA with annotations. The files are grouped into ''divisions'' that roughly correspond to taxonomic divisions. Some divisions are due to specific initiatives in biology and represent the functional categorization. There are currently 16 subdivisions. The data

in GenBank comes from two sources, direct submissions from the authors to one of the databases and bulk submissions from the sequencing centers in the form of ESTs, STS, GSS or large genomic records (usually sequences from cosmids, BACs or YACs). Data is exchanged daily with DDBJ, in Japan and EMBL, in Europe. They represent different distribution points, for the same information, in different formats. The US Office of Patent and Trademarks also contributes sequence data from issued patents. The GenBank, EMBL, and DDBJ nucleic acid sequence data banks have from their inception used tables of sites and features to describe the roles and locations of higher order sequence domains and elements within the genome of an organism.

Description of GenBank Flatfile (GBFF)

The GenBank Flatfile (GBFF) is the unit of information in the GenBank database. Each GBFF contains three parts. The **header** contains the descriptors (information) that apply to the whole record. The second part comprises the **features** that describe the annotations on the record and the third is the nucleotide **sequence** itself. All nucleotide database records end with "//" on the last line of the record.

Header

The header is the most database specific part of the record and variations in this exist between different databases. The first element on this line is the LOCUS name. This term historically represented the genetic LOCUS that was the subject of the record. This element starts with an alphabet and its length cannot exceed 10 characters. Characters after the first alphabet can be numeric or alphabetic and all letters are uppercase. The next component on the locus line is the length of the sequence in base pairs. In practice GenBank and other databases seldom accept sequences less than 50

bp and more than 350k bp in length. The next item on the locus line is the molecule type. The ‘‘mol-type’’ usually is DNA or RNA. The types are DNA, RNA, tRNA, rRNA, mRNA, and represent the original biological molecule. The next component is the three-letter GenBank divisional code. It has either taxonomic inferences or other classification purposes. These classifications are arbitrary and less useful as the taxonomic information is better represented in the Organism lines and the Source features of the GenBank entry. The date on the locus line is the date the record was last made public or first made public. It can also be the date of sequence submission. GenBank makes no claim on the modification of such reported dates. The definition line (also called the ‘‘DE’’ line) is the line in the GenBank record that attempts to summarize the biology of the record. The Accession number (‘‘AC’’) is the third line type of the record and is the key to reference a record in the database. The AC number is the number that is cited in the original publication and is the unique permanent identification code for the sequence. The NID line represents the gi (geninfo identifier) number for the nucleotide sequence. A gi number represents a unique identifier associated with a unique sequence. If the sequence changes, the gi will change, but the accession number will stay the same. The SOURCE line has the common name of the organism and its scientific name and taxonomic classification. Each GenBank record must have at least one REFERENCE or a citation. This field gives scientific credit to the authors of the work and typically points to the literature that explains the background context for the submitted sequence. When a reference is published, usually a MEDLINE identifier will be present. The CITATION/JOURNAL is present in most GenBank records giving scientific credit to the people responsible for the work surrounding the submitted sequence. It usually includes the postal address of the first author or main laboratory where the work was done. The last part

in the header section is the COMMENTS (also called ‘‘descriptors’’) that refer to the whole record. This segment is optional.

Feature Table (FT)

The middle segment of the GBFF record is the Feature Table (‘‘FT’’). This is the most direct representation of the biological information in the record. This section describes some of the key GenBank that offer annotations on specific portions of the sequence. The overall goal of the feature table design is to provide an extensive vocabulary for describing features in the sequence in a flexible enough framework so that they can be easily manipulated on a computer. Typically, through its feature table, each entry holds information about transcription, splicing and translation associated signals.

Eukaryotic Gene Structure in Genbank Feature Table

The CDS feature is the instruction to the reader for joining the two sequences together or on making an amino acid sequence from the indicated coordinates and the inferred genetic code. The GBFF maps all features through a DNA sequence coordinate system whereby the amino acid position is inferred through those of the DNA coordinates. This feature also illustrates the database cross-reference (db_xref) qualifier that allows the databases to cross-reference the sequence in question to an external database with a specific identifier of that database.

The location contains at least one sequence location descriptor and may contain one or more operators with one or more sequence location descriptors. Base numbers refer to the numbering in the entry. This numbering which is not necessarily the same as the numbering scheme used in the published report cited, designates the first base (5' end)

of the presented sequence as base 1. Base locations beyond the range of the presented sequence may not be used in location descriptors. The location descriptor can be one of the following:

a. Single base number.

b. Site between two indicated base numbers.

c. Single base chosen from within a specified range of bases.

d. Base numbers delimiting a sequence span.

e. Remote entry identifier followed by a local location descriptor (i.e., a to d).

A site between two points (nucleotides), such as endonucleolytic cleavage site is indicated by listing the two points separated by a carat (^). The first base number indicates a single base chosen from a range or span of bases and the last base number of the range separated by a single period (e.g 12.21) indicates a single base taken between the indicated points. The ‘‘less-than’’ symbol (<) and the ‘‘greater-than’’ symbol (>) indicate that a range ‘‘end point’’ is beyond and does not include the specified known base number. The starting base number indicates sequence spans and the ending base number separated by two periods (e.g., ‘’34..456’’). The ‘‘<’’ and ‘‘>’’ symbols may be used with the starting and ending base numbers to indicate that an end point is beyond (and does not include) the specified base number. A single point chosen from a range of points uses the ‘‘x.y’’ format described above. Feature labels are used in location descriptors only when they are required to improve readability. A location in a remote entry (not the entry to which the feature table belongs) can be specified in either of two ways: by specifying the remote entry (by Accession number) followed by a location descriptor which applies to that entry's sequence, or by specifying the remote entry followed by the label of a feature in that

entry's feature table. The location operator is a prefix that specifies what must be done to the indicated sequence to find or construct the location corresponding to the feature. The allowable operators are:

(1) “complement (location)” means the complement of the presented sequence in the span specified by ‘‘location’’ (i.e., read the complement of the presented strand in its 5' to 3' direction).

(2) Join (location, location . location) implies that the indicated elements should be joined (placed end to end) to form one contiguous sequence.

(3) Order (location, location .. location) means that the elements can be found in the specified order (5' to 3' direction), but nothing is implied about the reasonableness about joining them.

SEQ

The “feature table” is followed by the SEQ line that gives information on the composition of the nucleotide sequence followed by the nucleotide sequence.

1.10 Genome Revolution and Eukaryotic Gene Structure

Advancements in genetic engineering, micro-fabrication techniques and nano-technology have revolutionized biotechnology in recent years (West and Halas, 2000). The huge data generated by rapid genome sequencing has been subjected to data-warehousing, clustering, curation and by post-genome analysis. This has resulted in several specialized databases (Baxevanis, 2000). GenBank, the main repository of nucleotide sequence data contains more than 3.4 billion nucleotide bases from over 55, 000 different organisms and doubles every 15 months (Benson *et al.,* 2000). Based on the available knowledge in the GenBank feature table, the accuracy of

computational gene-finding algorithms has improved significantly over the past years (Stormo, 2000; Burge and Karlin, 1998; Burge and Karlin, 1997; Wu, 1996). A normal mathematical formalism is often inadequate to describe gene structures in eukaryotic genes. A better description of exon-intron organization in eukaryotic genes will aid in tracing the connecting control mechanisms of their functions in protein products. A number of authors have conducted comprehensive analysis on exon-intron distribution in eukaryotic genes (Deutsch and Long, 1999; Tomita *et al.,* 1996; Hawkins, 1988). However, little is known about the principles governing exon intron organization in eukaryotic genomes. In this thesis we investigate three aspects of eukaryotic genomes:

(1) To investigate the length distribution properties for exons & introns in 6 eukaryotic genomes.

(2) To investigate the characteristics of FDA approved drugs & the role of SEG as drug targets.

(3) Create a DB on mismatched introns from human and mouse.

1.11 Organization of The Thesis

Corresponding to the above 3 objectives, the following chapters in this thesis are:

- Chapter 2: Intron & exon length distribution for 6 eukaryotic genomes. To make use of the completed sequenced 6 eukaryotic genomes available on GenBank, their intron, exon and intergenic length information is extracted, and the length distribution patterns of intron and exon are plotted, so that some regularity and similarity can be identified at the genome structure level.
- Chapter 3: Characteristics of targets for FDA approved drugs. A detailed study on the five genome structure level characteristics of FDA approved drugs,

which are the number of pathway, number of tissue, number of protein-protein interaction, number of protein families and number of exons, is reported.

- Chapter 4: Database on mismatched introns (MIDB). In order to build a platform for the ease of research on mismatch intron, it is desired to have the intron and exon information for the orthologues and homologues from 2 or more species. In this research, firstly homologous and orthologous genes for human and mouse are identified, and then the introns and exons information is extracted out from the NCBI data file based on the homologue and orthologue pairs. Users can analyze the intron mismatch pattern of human mouse homologue or orthologue with all the above data accessible on web.
- Chapter 5: Conclusion and future work.

CHAPTER 2 INTRON, EXON LENGTH DISTRIBUTIONS FOR 6 EUKARYOTIC GENOMES

2.1 Introduction

The availability of complete genome sequences for many eukaryotic organisms continues to contribute towards a better understanding of their genome design and evolution. Vertebrate genes are typically split into numerous small exons interrupted by much larger introns (10 or 100 times longer) (Hawkins, 1988). Exon-intron architecture varies across the eukaryotic kingdom with genes with small exons usual in vertebrates and genes with small introns being normal in invertebrates (Sterner *et al.,* 1996). A number of available informational systems on various gene characteristics complement each other and are indispensable for many genomic studies. For example, EID, Entrez Gene, Ensembl Genome Browser, UCSC Genome Browser, SpliceNest, Xpro, ISIS, ExInt and others, a fraction of which can be found in Galperin's biological database repository. These databases present diverse information on the genes and also provide web-based interfaces for quick and simple access to the data.

With the tremendous increase in genome sequencing projects, a variety of computational techniques have been developed to in a indirect manner infer gene structure from genome sequence data, including the detection of intron–exon boundaries; see Zhang (2002) for a comprehensive review. These techniques have been proved to perform better such that they can recognize the large number of intron–exon boundaries within the CDS. Analysis on exon–intron gene structures is complex and non-trivial due to enormous expansions of the eukaryotic genomes

(Gregory *et al.* 2007), great variety of gene forms, and the imperfectness in sequence data. Thus, gene annotations constantly expand and update with refinement in gene annotation programs creating the need for constant explorations on genome architecture and design in the expanding variety of completely sequenced eukaryotic genomes and their analyses. Examination of intron and exon characteristics can reveal the nature of the underlying mechanism that recognises the intron in order for it to be spliced. These analyses also help identify the patterns and regularities in eukaryotic genomes.

With the purpose to understand how the genes characteristics of the exon-intron structures evolved, the first collection of exon-intron structures in eukaryotic genes was published by *Hawkins in 1988*. Since then many disparate reports have been presented and the use of patterns in exons and introns to understand gene structure is becoming increasingly ubiquitous. Also, the sequential arrangement of coding (exons) and non-coding (introns) regions is of particular interest from a biological viewpoint in revealing essential details necessary for understanding the assembly of the spliceosome and the splicing process in general. Deutsch *et al.* reported exon-intron structures from eukaryotic model organisms and analyzed the statistical distribution of spliceosomal introns (splicing of these introns requires the participation of a specific set of protein-RNA particles) and exons of nuclear genes in 10 eukaryotic model organisms from GenBank (Deutsch and Long, 1999). Concurrently, Sakharkar et al. provided a distribution of genes, exons and introns in the human and mouse genomes and discerned correlations between them (Sakharkar *et al,* 2004; Sakharkar *et al.*, 2005). The analyses provide a general picture of gene architecture of intron-containing human and mouse genes. The results suggested that the total length in introns and

intergenic DNA on each chromosome is significantly correlated to the determined chromosome size (genome size) and provided insight to their role in shaping and structuring of both the human and mouse genomes (Sakharkar *et al.*, 2005).

In this manuscript, we present exon-intron length profile analyses for six crown eukaryotic genomes - human, mouse, chimpanzee, zebrafish, worm and fly. A statistical analysis on length comparisons of first exon and first intron, and last exon and last intron with the average length of remaining exon or intron is also provided. These analyses provide a quantitative and qualitative view of genome organization and the findings could help improve gene structure prediction by computational methods.

2.2 Materials and Methods

The *H.sapiens* genome data (Aug 2006), the *P .troglodytes* genome data (May 2007), the *M.musculus* genome data (Mar 2007), , the *D.rerio* genome data (Apr 2007), the *C.elegans* genome data (Apr 2007) and the *D. melanogaster* genome data (Jul 2007), were downloaded from the National Center for Biotechnology Information (NCBI) at ftp://ftp.ncbi.nlm.nih.gov/genomes/. The data were processed for extraction of exons and introns based on the CDS feature table annotation, as described elsewhere (Sakharkar *et al.*, 2002).

Starting with 32930 CDS in *H.sapiens*, 51259 CDS in *P.troglodytes*, 27406 CDS in *M.musculus*, 32219 CDS in *D.rerio*, 22844 CDS in *C.elegans* and 19765 CDS in *D.melanogaster*, we filtered out 280397 exons and 248031 introns for *H.sapiens*, 478532 exons and 427280 introns for *P.troglodytes*, 226677 exons and 199282 introns

for *M.musculus*, 255249 exons and 223049 introns for *D.rerio*, 145608 exons and 122764 introns for *C.elegans* and 89336 exons and 69578 introns for *D.melanogaster* [Table 2.1-2.6]. Results of exons (exons in the coding region) and introns (introns between the coding exons) length distributions were tabulated for further analysis and correlations among them were discerned. For the identification of first and last exons and introns, the data was extracted using the mRNA feature and its comparison with the CDS.

Chr #	#introns	# of exons	# of CDS	Exon/CDS	Intron	Exon	Intergenic	Chr size (□ locus bp)	Length/Chr size Intron	Length/Chr size Exon	Length/Chr size Igd	Shortest		Longest		Average Length		Stand deviation	
												Exon	Intron	Exon	Intron	Exon	Intron	Exon	Intron
1	28933	32359	3426	9.45	117208860	5223531	141537920	226212984	51.80%	2.30%	62.60%	1	1	12048	397263	161.4	4051.0	223.3	12205.1
2	18807	20891	2162	9.66	107288794	3593372	160039477	237898220	45.10%	1.50%	67.30%	1	30	17106	483412	170.8	5703.8	346.2	17247.7
3	14172	15883	1711	9.28	91800662	2577343	129677606	195304083	47.00%	1.30%	66.40%	1	30	6654	522714	162.3	6477.6	217.9	21301.5
4	8947	10162	1215	8.36	66587808	1787338	131983125	187939711	35.40%	1.00%	70.20%	1	37	6255	494708	175.9	7442.5	256.6	21096.5
5	10518	12022	1504	7.99	72961432	2202695	122671223	177846453	41.00%	1.20%	69.00%	1	32	10489	370360	183.2	6936.8	331.7	20265.2
6	11289	12833	1544	8.31	69625838	2231284	114570258	169099554	41.20%	1.30%	67.80%	1	30	7152	469892	173.9	6167.6	256.8	18857.6
7	12152	13799	1749	7.89	85780039	2335430	99459883	155402083	55.20%	1.50%	64.00%	1	30	12219	657297	167.1	7056.8	262.0	23455.9
8	8282	9467	1185	7.99	57149089	1659136	98981018	143330736	39.90%	1.20%	69.10%	1	53	7263	387853	175.3	6900.4	316.7	19628.5
9	10195	11574	1387	8.34	54635136	1933732	78979826	120989686	45.20%	1.60%	65.30%	1	33	6436	256666	166.8	5358.9	248.4	13660.3
10	11294	12674	1380	9.18	74656963	1964101	81327014	131738012	56.70%	1.50%	61.70%	1	30	7812	482575	155.0	6610.3	228.9	20034.0
11	12655	14517	1862	7.8	56854453	2551815	82549621	131246147	43.30%	1.90%	62.90%	1	31	17331	589253	175.8	4492.6	274.7	16627.1
12	13238	14800	1562	9.48	58143236	2319730	85774199	130303534	44.60%	1.80%	65.80%	1	30	5002	328545	156.7	4392.1	178.6	12582.8
13	3843	4368	525	8.32	32048851	786128	68459745	95746838	33.50%	0.80%	71.50%	1	33	11555	740920	180.0	8339.5	343.9	30462.6
14	7045	7957	1205	6.6	40038526	1460988	60517862	88290585	45.30%	1.70%	68.50%	1	43	8340	479079	173.5	5677.7	254.3	19006.3
15	9202	10358	1156	8.96	43294501	1682291	52139185	81926261	52.80%	2.10%	63.60%	2	1	8035	550366	162.4	4704.9	236.9	13082.9
16	12410	13890	1480	9.39	36696726	2391835	52653803	78990239	46.50%	3.00%	66.70%	1	30	8612	466049	172.2	2957.0	239.2	10685.7
17	14583	16279	1696	9.6	44268755	2650124	45295326	79617833	55.60%	3.30%	56.90%	1	31	5565	292213	162.8	3035.6	204.6	9486.2
18	3478	3952	474	8.34	29310923	685131	51298510	74660417	39.30%	0.90%	68.70%	1	32	4721	411175	173.4	8427.5	248.3	21130.0
19	12225	14102	1877	7.51	24447601	2597142	33439133	56037509	43.60%	4.60%	59.70%	1	1	5059	96613	184.2	1999.8	273.2	4297.0
20	6216	7077	861	8.22	30076918	1131615	40706586	59505253	50.50%	1.90%	68.40%	3	60	3738	544980	159.9	4838.6	213.0	17892.0
21	3216	3675	459	8.01	16866227	588864	25332217	35451691	47.60%	1.70%	71.50%	1	64	5916	323563	160.2	5244.5	266.7	15605.3
22	5497	6304	890	7.08	20526198	1051942	21375358	35058650	58.50%	3.00%	61.00%	1	34	6762	268179	163.2	3734.1	255.8	10510.4
ChrX	9071	10538	1467	7.18	63085108	1938478	115297988	152577922	41.30%	1.30%	75.60%	1	33	6042	536480	184.0	6954.6	289.6	22389.1
ChrY	763	916	153	5.99	4180725	153201	19320069	25652954	16.30%	0.60%	75.30%	3	64	2493	400349	167.3	5479.3	234.8	19119.1
Total	248031	280397	32930	8.5	1297533369	47497246	1913386952	2870827355											

Table 2.1
Summary of *H.sapiens* Genome data by Chromosome.

Chr	#introns	# of exons	# of CDS	Exon/CDS	Intron	Exon	Intergenic	Chr size (∑ locus bp)	Length/Chr size Intron	Length/Chr size Exon	Length/Chr size Igd	Shortest		Longest		Average Length		Stand deviation	
												Exon	Intron	Exon	Intron	Exon	Intron	Exon	Intron
1	43786	49165	5379	9.14	205606594	7827255	130705253	232780216	88.30%	3.40%	56.10%	1	30	9129	452683	159.2	4695.7	199.2	13562.3
2a	16039	17900	1863	9.61	90209332	2758669	74586326	111497100	80.90%	2.50%	66.90%	1	30	7572	492350	154.1	5624.1	192.2	15445.7
2b	19797	21677	1880	11.53	118304437	3268304	83645109	132835831	89.10%	2.50%	63.00%	1	32	6877	423212	150.8	5975.9	194.1	18265.3
3	28627	31679	3052	10.38	204781594	4927802	127848229	202464459	101.10%	2.40%	63.10%	1	30	6137	503647	155.6	7153.4	195.1	23128.4
4	19617	21926	2309	9.50	150456434	3550176	129082333	197326094	76.20%	1.80%	65.40%	1	33	6255	535821	161.9	7669.7	216.7	20970.8
5	21390	23922	2532	9.45	142889600	3964393	120888845	182067534	78.50%	2.20%	66.40%	1	30	6577	689076	165.7	6680.2	273.4	19286.6
6	20057	22623	2566	8.82	120939256	3731329	114835023	177555873	68.10%	2.10%	64.70%	1	30	7152	475829	164.9	6029.8	218.2	18362.6
7	20792	23338	2546	9.17	155992643	3639335	92663305	162359053	96.10%	2.20%	57.10%	1	30	4779	672886	155.9	7502.5	208.3	22580.9
8	15469	17364	1895	9.16	111212120	2617309	93109683	148638763	74.80%	1.80%	62.60%	1	30	7188	698552	150.7	7189.4	187.9	19340.7
9	17524	19533	2009	9.72	89867562	3119658	77630081	120061799	74.90%	2.60%	64.70%	1	30	6598	259262	159.7	5128.3	225.1	13103.4
10	19662	21918	2256	9.72	132125597	3249484	84586724	137441083	96.10%	2.40%	61.50%	1	30	6848	522632	148.3	6719.8	198.4	20775.6
11	26748	29963	3215	9.32	126453146	4797224	84217203	135429951	93.40%	3.50%	62.20%	1	30	7206	668104	160.1	4727.6	205.9	17824.3
12	22399	24951	2552	9.78	102775710	3814419	88956566	135675203	75.80%	2.80%	65.60%	1	30	3818	316780	152.9	4588.4	164.0	11913.4
13	9062	10033	971	10.33	63909190	1547131	59494587	98704794	64.70%	1.60%	60.30%	1	30	11555	710571	154.2	7052.4	274.2	22946.4
14	14487	16222	1739	9.33	97042281	2592994	60903847	90582208	107.10%	2.90%	67.20%	1	31	6828	485763	159.7	6698.1	191.7	24317.9
15	16482	18225	1743	10.46	79383931	2851851	48573826	82071288	96.70%	3.50%	59.20%	1	30	9530	561408	156.5	4816.4	202.4	12204.0
16	16481	18576	2096	8.86	64165847	2906227	43160909	83696349	76.70%	3.50%	51.60%	1	31	5493	508707	156.4	3893.2	182.3	13574.0
17	23878	26711	2833	9.43	76511070	4130027	47935138	81665014	93.70%	5.10%	58.70%	1	30	5550	297336	154.6	3204.2	200.2	9152.3
18	7094	7939	845	9.40	55204356	1259103	52504578	77548041	71.20%	1.60%	67.70%	1	35	4764	420394	158.6	7781.8	242.0	17037.2
19	16521	19029	2508	7.59	35552804	3356530	34812892	58176543	61.10%	5.80%	59.80%	1	30	5058	291065	176.4	2152.0	254.9	5841.8
20	11287	12638	1351	9.35	49116180	1897460	41322261	61944263	79.30%	3.10%	66.70%	2	30	3108	547962	150.1	4351.6	188.2	14214.1
21	5584	6219	635	9.79	27733976	976297	23547356	32724799	84.70%	3.00%	72.00%	1	40	5916	318495	157.0	4966.7	271.5	14874.8
22	8262	9348	1086	8.61	27439990	1450229	18671594	35163897	78.00%	4.10%	53.10%	1	30	5661	333697	155.1	3321.2	197.6	9523.2
ChrX	5324	6574	1250	5.26	33413780	1283549	63420556	150212081	22.20%	0.90%	42.20%	1	30	4719	486283	195.2	6276.1	280.6	20625.6
ChrY	911	1059	148	7.16	4045302	157018	13897685	23467553	17.20%	0.70%	59.20%	3	60	1595	100054	148.3	4440.5	169.9	12716.2
Total	427280	478532	51259	9.3	2365132732	75673773	1810999909	2952089789											

Table 2.2
Summary of *P.troglodytes* Genome data by Chromosome.

Chr #	#introns	# of exons	# of CDS	Exon/CDS	Intron	Exon	Intergenic	Chr size (□ locus bp)	Length/Chr size Intron	Length/Chr size Exon	Length/Chr size Igd	Shortest		Longest		Average Length		Stand deviation	
												Exon	Intron	Exon	Intron	Exon	Intron	Exon	Intron
1	13233	14808	1575	9.40	72548755	2573960	134101939	191534870	37.90%	1.30%	70.00%	1	30	8558	396454	173.8	5482.4	257.0	16393.8
2	17530	19766	2236	8.84	67157390	3567825	123958612	178726762	37.60%	2.00%	69.40%	1	32	17106	362815	180.5	3831.0	317.4	12905.1
3	9269	10549	1280	8.24	48575176	1817729	116397606	156457112	31.00%	1.20%	74.40%	1	1	7993	472564	172.3	5240.6	250.8	18046.1
4	11599	13221	1622	8.15	47515047	2319608	105358843	151489701	31.40%	1.50%	69.50%	1	30	8020	375391	175.4	4096.5	254.5	12139.6
5	13417	15052	1635	9.21	56963253	2590451	102115895	150168310	37.90%	1.70%	68.00%	1	30	5953	564636	172.1	4245.6	233.2	14282.3
6	9747	11205	1458	7.69	53633615	1970225	98992329	146293685	36.70%	1.30%	67.70%	1	31	4740	593603	175.8	5502.6	223.8	20193.0
7	13927	16336	2409	6.78	47390134	3411487	98675951	141373004	33.50%	2.40%	69.80%	1	30	7293	412932	208.8	3402.8	320.5	11549.5
8	10622	11940	1318	9.06	45936373	2047875	87545615	125472059	36.60%	1.60%	69.80%	1	33	7225	639386	171.5	4324.6	281.7	16206.5
9	11558	13015	1457	8.93	46887361	2288023	77482838	120717901	38.80%	1.90%	64.20%	1	1	11544	612881	175.8	4056.7	276.0	15439.1
10	9251	10514	1263	8.32	50718680	1879282	88960372	126819929	40.00%	1.50%	70.10%	1	1	5139	487269	178.7	5482.5	236.2	19974.3
11	14967	16896	1929	8.76	46288759	2897009	76996710	118711832	39.00%	2.40%	64.90%	1	42	7442	479339	171.5	3092.7	233.1	10500.1
12	7253	8261	1008	8.20	36057275	1452633	85844326	117082159	30.80%	1.20%	73.30%	1	39	7467	429635	175.8	4971.4	265.5	15810.1
13	6448	7500	1052	7.13	39003805	1432579	81440242	117166569	33.30%	1.20%	69.50%	2	35	6490	707703	191.0	6049.0	316.7	19515.0
14	7944	9071	1127	8.05	43721282	1536488	86442187	120927170	36.20%	1.30%	71.50%	1	30	4805	735947	169.4	5503.7	221.6	24867.4
15	8699	9698	999	9.71	35791401	1755527	71905245	100493509	35.60%	1.70%	71.60%	1	30	10318	275464	181.0	4114.4	339.6	13083.9
16	6249	7093	855	8.30	33914154	1223049	67633140	95229459	35.60%	1.30%	71.00%	1	32	6410	666870	172.1	5427.1	259.0	18704.9
17	9075	10362	1287	8.05	32650027	1907905	64454451	92111511	35.40%	2.10%	70.00%	1	30	6569	277725	184.1	3597.8	255.0	12955.5
18	4782	5459	677	8.06	32114681	1092181	63923079	87565837	36.70%	1.20%	73.00%	2	30	6577	395281	200.1	6715.7	344.0	19254.9
19	6740	7611	871	8.74	30028226	1304219	35913490	58121190	51.70%	2.20%	61.80%	1	43	16626	558120	171.4	4455.2	306.5	20919.4
ChrX	6663	7929	1266	6.26	34777653	1583001	127530733	160546165	21.70%	1.00%	79.40%	1	30	4993	522727	199.6	5219.5	323.9	17846.0
ChrY	309	391	82	4.77	944510	73636	4613013	15707136	6.00%	0.50%	29.40%	2	64	2337	26978	188.3	3056.7	265.4	4300.0
Total	199282	226677	27406	8.27	902617557	40724692	1800286616	2572715870											

Table 2.3
Summary of *M.musculus* Genome data by Chromosome.

Chr #	#introns	# of exons	# of CDS	Exon/CDS	Intron	Exon	Intergenic	Chr size (□ locus bp)	Length/Chr size Intron	Length/Chr size Exon	Length/Chr size Igd	Shortest		Longest		Average Length		Stand deviation	
												Exon	Intron	Exon	Intron	Exon	Intron	Exon	Intron
1	11295	12835	1540	8.33	30012876	2263396	34920208	70571795	42.50%	3.20%	49.50%	1	30	6870	846156	176.3	2657.2	272.5	13293.6
2	10574	12078	1508	8.01	26684330	2122008	29795363	61873085	43.10%	3.40%	48.20%	1	30	6894	736842	175.5	2521.1	264.2	9796.4
3	10632	12328	1706	7.23	28984342	2244250	38610394	77158595	37.60%	2.90%	50.00%	1	30	6081	469269	181.7	2725.6	287.7	9700.0
4	6902	7907	1005	7.87	17173929	1454281	25978304	47238102	36.40%	3.10%	55.00%	1	30	6166	214921	183.9	2488.3	293.5	7759.7
5	12868	14601	1734	8.42	35274271	2396000	39357850	84631280	41.70%	2.80%	46.50%	1	30	7791	496139	164.1	2741.2	235.6	9428.4
6	10934	12368	1434	8.62	28005773	2053869	32233089	69534419	40.30%	3.00%	46.40%	1	30	9078	635566	166.1	2561.3	263.7	11323.3
7	11658	13413	1755	7.64	35112201	2357146	40703097	87668671	40.10%	2.70%	46.40%	1	30	8352	862314	175.7	3011.9	283.1	15142.0
8	10756	12255	1499	8.18	30785335	2005024	30951327	66781301	46.10%	3.00%	46.30%	1	31	5214	483431	163.6	2862.2	222.1	11058.1
9	8184	9256	1072	8.63	23296445	1619321	28179428	55700484	41.80%	2.90%	50.60%	1	32	11961	631298	174.9	2846.6	309.7	11377.8
10	8005	9181	1176	7.81	23375534	1694211	23949030	54055995	43.20%	3.10%	44.30%	1	30	6325	491652	184.5	2920.1	305.3	11493.8
11	6743	7766	1023	7.59	22716453	1356578	25628807	52330080	43.40%	2.60%	49.00%	1	30	9849	613910	174.7	3368.9	282.9	13313.8
12	8975	10229	1256	8.14	23407452	1736361	26534535	58699258	39.90%	3.00%	45.20%	1	30	8979	934518	169.6	2607.3	249.1	12517.6
13	9263	10504	1241	8.46	27158406	1812318	28801114	64241675	42.30%	2.80%	44.80%	1	32	6852	326631	172.5	2931.9	277.1	8967.0
14	9127	10695	1568	6.82	29497055	2140745	49235686	91695635	32.20%	2.30%	53.70%	1	30	10026	760951	200.2	3231.8	348.2	12156.9
15	8521	9748	1227	7.94	23330659	1755646	26585490	57197518	40.80%	3.10%	46.50%	1	30	6810	311339	180.1	2738.0	281.0	8880.3
16	9225	10633	1408	7.55	27482205	1840949	31158610	65471547	42.00%	2.80%	47.60%	1	30	4985	467174	173.1	2979.1	244.3	10282.7
17	9247	10482	1235	8.49	26187428	1884080	28228986	63392620	41.30%	3.00%	44.50%	1	34	8852	414647	179.7	2832.0	296.1	9471.8
18	7331	8347	1016	8.22	23123834	1463460	24620266	59745843	38.70%	2.40%	41.20%	1	30	8455	241507	175.3	3154.3	258.7	9828.4
19	6765	7798	1033	7.55	19569802	1451910	26804916	51704704	37.80%	2.80%	51.80%	1	30	11853	246811	186.2	2892.8	381.2	8426.2
20	9800	11143	1343	8.30	25666395	2021473	30689450	63637807	40.30%	3.20%	48.20%	1	30	6963	291483	181.4	2619.0	280.6	6984.8
21	8132	9386	1254	7.48	23078615	1593198	25328269	56235177	41.00%	2.80%	45.00%	1	30	11253	821466	169.7	2838.0	280.3	12669.1
22	8268	9574	1306	7.33	17504978	1720716	22659343	47736866	36.70%	3.60%	47.50%	1	1	6985	177653	179.7	2117.2	257.4	5623.1
23	7931	9014	1083	8.32	19882775	1504980	24733607	53198997	37.40%	2.82%	46.50%	2	30	4470	272813	167.0	2507.0	223.4	7129.1
24	5649	6479	831	7.80	19174928	1081636	23632653	46068729	41.60%	2.30%	51.30%	3	32	6096	672677	166.9	3394.4	246.3	16910.9
25	6264	7229	966	7.48	16388590	1177094	18783935	40299440	40.70%	2.90%	46.60%	1	33	10288	885678	162.8	2615.8	261.0	13358.6
Total	223049	255249	32219	7.92	622874611	44750650	738103757	1546869623											

Table 2.4
Summary of *D.rerio* Genome data by Chromosome.

Chr #	#introns	# of exons	# of CDS	Exon/CDS	Intron	Exon	Intergenic	Chr size (□ locus bp)	Length/Chr size Intron	Length/Chr size Exon	Length/Chr size Igd	Shortest		Longest		Average Length		Stand deviation	
												Exon	Intron	Exon	Intron	Exon	Intron	Exon	Intron
1	19356	22684	3328	6.82	7604686	4890615	6279384	15072418	50.50%	32.40%	41.70%	1	31	11688	21230	215.6	392.9	282.2	883.2
2	19707	23677	3970	5.96	5947973	5062329	7127653	15279313	38.90%	33.10%	46.60%	1	32	8069	20124	213.8	301.8	223.5	764.3
3	17382	20551	3169	6.49	6525148	4429992	5840805	13783317	47.30%	32.10%	42.40%	1	36	7204	20573	215.6	375.4	235.7	793.8
4	19286	22994	3708	6.20	6525745	4724712	9060154	17493785	37.30%	27.00%	51.80%	1	31	14975	20948	205.5	338.4	241.0	719.8
5	25235	30628	5393	5.68	6563505	6609597	10004911	20922233	31.40%	31.60%	47.80%	2	34	12217	15573	215.8	260.1	283.7	559.4
X	21798	25074	3276	7.65	6119301	4482627	9998044	17718851	34.50%	25.30%	56.40%	2	30	10633	20393	178.8	280.7	199.6	689.0
Total	122764	145608	22844	6.30	39286358	30199872	48310951	100269917											

Table 2.5
Summary of *C.elegans* Genome data by Chromosome.

Chr #	#introns	# exons	# CDS	#Exon/CDS	Intron	Exon	Intergenic	Chr size (□ locus bp)	Intron	Exon	Intergenic	Shortest		Longest		Average Length		Stand deviation	
												Exon	Intron	Exon	Intron	Exon	Intron	Exon	Intron
2	27941	35607	7666	4.64	27112353	13218797	27449264	44158252	61.4%	29.9%	62.2%	1	1	27708	108407	371.2	970.3	524.7	3693.0
3	29236	37870	8641	4.38	31344361	14271662	67625116	52448610	59.8%	27.2%	128.9%	1	32	27591	132737	376.9	1072.1	513.8	4594.2
4	1665	1885	220	8.36	1602054	642123	607067	1351857	118.5%	47.5%	44.9%	2	47	9471	20173	340.6	962.2	645.2	1968.7
X	10736	13974	3238	4.32	11063783	5702107	14607230	22422827	49.3%	25.4%	65.1%	1	11	14581	113165	408.1	1030.5	566.1	4080.3
Total	69578	89336	19765	4.5	71122551	33834689	110288677	120381546											

Table 2.6
Summary of *D. Melanogaster* Genome data by Chromosome.

2.3 Result and Discussion

The distributions for the number of CDS, the exons and introns and correlations between them are presented for the genomes of *H.sapiens, P.troglodytes, M.musculus, D.rerio, C.elegans* and *D.melanogaster* (Table 2.1 to 2.6).

Deutsch et al., reported that on average a human gene contains 5 exons per gene and a mouse gene contains on average 4.4 exons per gene (Deutsch and Long, 1999). This analysis was performed using GenBank data. The genome data is non-redundant and circumvents the technical challenge in data purging. Earlier, we investigated on the exon-intron distribution profiles for the intron-containing genes in the human and mouse genomes (Sakharkar *et al.,* 2004; Sakharkar *et al,* 2005). However, recently databases have evolved both in content and size during and the genome sequence for several eukaryotic genomes is now available. The distributions for the number of genes against the number of exons in *H.sapiens, P.troglodytes, M.musculus, D.rerio, C.elegans* and *D.melanogaster* chromosomes are presented [Figure 2.1].

2.3.1 Chromosome Size and Archtecture

The largest number of bases sequenced (∑ locus size) is for chromosome 1 and chromosome 2 and smallest is for chromosome Y in *H. sapiens*; the largest number of bases sequenced (∑ locus size) is for chromosome 1 and smallest is for chromosome Y in *P.troglodytes*; the largest number of bases sequenced (∑ locus size) is for chromosome 1 and smallest is for chromosome Y in *M.musculus*; The largest number of bases sequenced (∑ locus size) is for chromosome 14 and smallest is for chromosome 25 in *D. rerio*; The largest number of bases sequenced (∑ locus size) is

for chromosome 5 and smallest is for chromosome 3 in *C.elegans*; and the largest number of bases sequenced (∑ locus size) is for chromosome 3 and smallest is for chromosome 4 in *D.melanogaster*.

2.3.2 Genes and Gene Density

Chromosome 1 (3426) has the largest number of CDS and chromosome Y (153) has the smallest number of CDS in *H.sapiens* and in *P. troglodytes* (148) (5379), chromosome 7 (2409) has the largest number of CDS and chromosome Y (82) has the smallest number of CDS in *M.musculus*, Chromosome number 7 (1755) has the largest number of CDS and chromosome 24 (831) has the smallest number of CDS in *D.rerio*, Chromosome number 5 (5393) has the largest number of CDS and chromosome 3 (3169) has the smallest number of CDS in *C.elegans* and Chromosome number 3 has the largest number of CDS (8641) and chromosome 4 (220) has the smallest number of CDS in *D. melanogaster*.

The gene density (genes/Mb) is lowest for chromosome 13 (5.48) and highest for chromosome 19 (33.49) in *H. sapiens*; the gene density (genes/Mb) is lowest for chromosome Y (6.30) and highest for chromosome 19 (43.11) in *P. troglodytes*; the gene density (genes/Mb) is lowest for chromosome Y (5.22) and highest for chromosome 7 (17.04) in *M.musculus*; the gene density (genes/Mb) is lowest for chromosome 18 (17.00) and highest for chromosome 22 (27.35) in *D.rerio*; the gene density (genes/Mb) is lowest for chromosome X (184.88) and highest for chromosome 2 (259.82) in *C.elegans*; and the gene density (genes/Mb) is lowest for chromosome X (144.40) and highest for chromosome 2 (173.60) in *D. melanogaster*. Overall it appears that gene density is lowest for *M .musculus* (10.65), and increases from *H.*

sapiens (11.47), *P.troglodytes* (17.36), *D. rerio* (20.82), *D. melanogaster* (164.18), *C. elegans* (227.82).

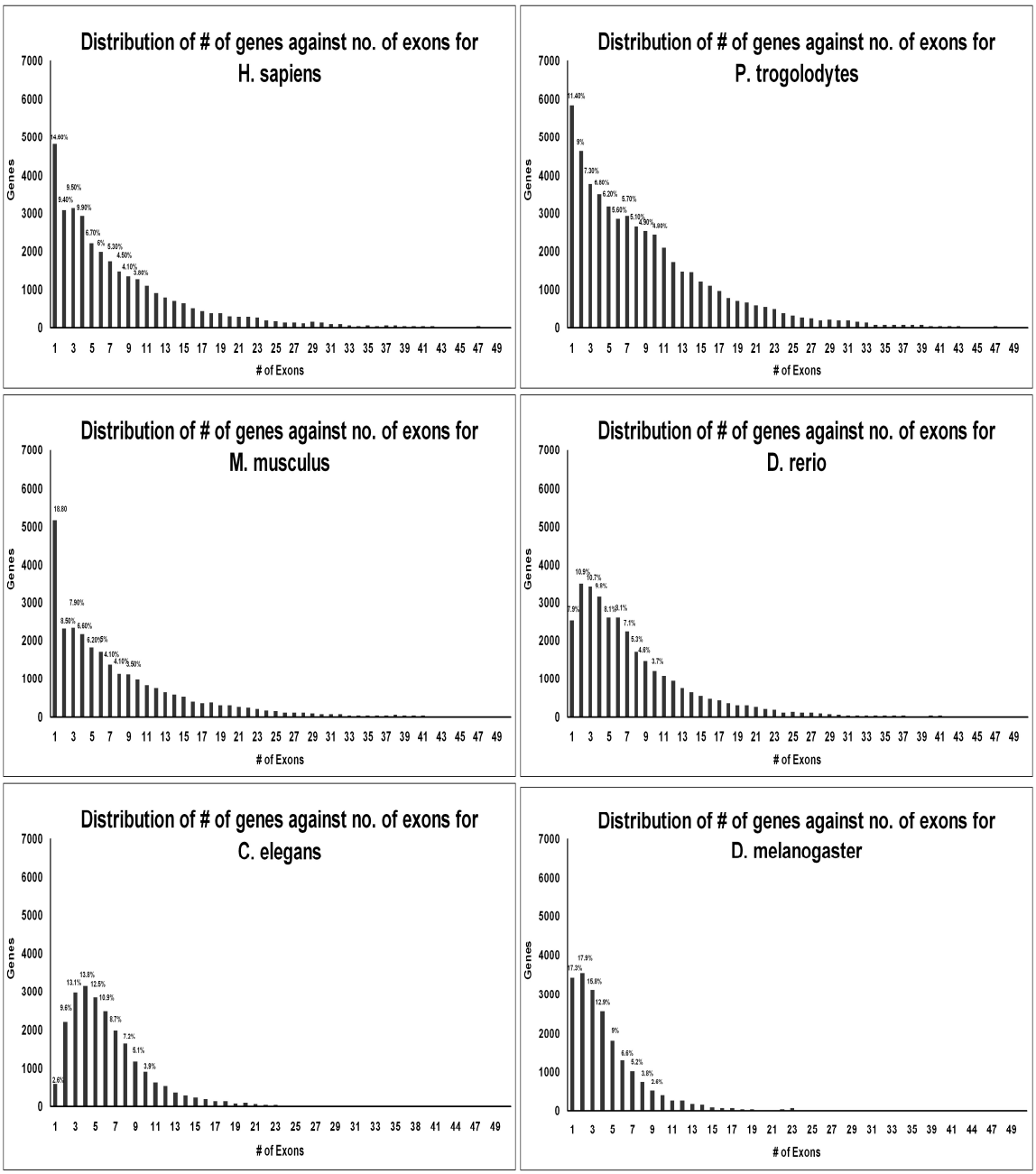

Figure 2.1
Distribution # of genes against # of exons for the six genomes.

2.3.3 Exons and Intron Distribution

Exonic DNA occupies ~1.65% of genome space in *H. sapiens*, ~2.56% in *P. troglodytes*; ~1.58% in *M.musculus*, ~2.89% in *D.rerio* and ~30.12% in *C.elegans* and ~28.11% in *D. melanogaster* genome. Introns and intergenic DNA occupy the remaining of the genome space, ~45.20%, ~66.65%; ~80.12%, ~61.35%; ~35.08, ~69.98; ~40.27, ~47.72; ~39.18%, ~48.18%; and ~59.08%, ~91.62% in H. *sapiens*, *P. troglodytes, M.musculus, D.rerio, C.elegans* and *D. melanogaster*, respectively [Figure 2.2]. It must be noted that the total exceeds 100%, this may be due to re-sequencing segments in consecutive locus regions and may also be due to the presence of overlapping genes in eukaryotic genomes. Nonetheless, the data supports previous observations and the fact that only a fraction of eukaryotic genome is coding.

It is observed that the average number of exons per gene (CDS) range from 5.99 (Chromosome Y) to 9.66 (Chromosome 2) in *H. sapiens*; 5.26 (Chromosome X) to 11.53 (Chromosome 2b) in *P. troglodytes*; 4.77 (Chromosome Y) to 9.71 (Chromosome 15) in *M.musculus*; 6.82 (Chromosome 14) to 8.63 (Chromosome 9) in *D. rerio*; 5.67 (Chromosome 5) to 7.68 (Chromosome X) in *C. elegans* and 4.32 (X) to 8.36 (Chromosome 4) in *D. melanogaster* genome. The average number of exons in the genome is 8.5/gene in *H. sapiens*; 9.3/gene in *P. troglodytes*; 8.27/gene in *M.musculus*; 7.92/gene in *D.rerio*; 6.3/gene in *C.elegans* and 4.5/gene in *D. melanogaster* [Table 2.1 – 2.6]. 80% of the genes had less than 15 exons in all the six genomes under investigation [Figure 2.2]. In general the maximum number of exons and introns are found on chromosomes with maximum number of CDS. These results also support the fact that longer proteins in general are encoded by genes with more number of exons that is they are more split.

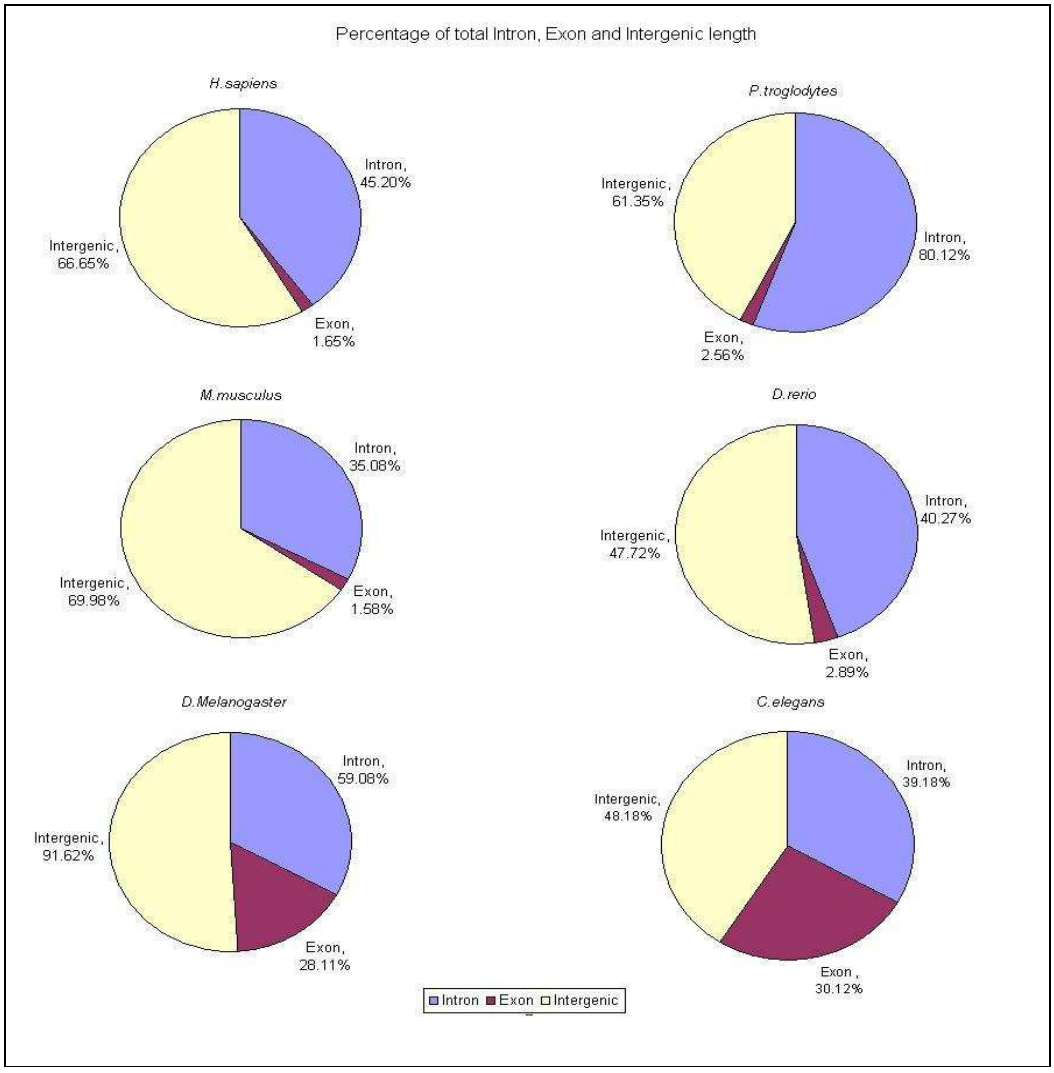

Figure 2.2
Exon, Intron and Intergenic DNA distribution in % for the six genomes

2.3.4 Exon and Intron Length Distribution

Ever since the discovery of introns, there has been intense debate about their origins, stability, and adaptive significance. The results show that exon length are distributed much more tightly than introns for all the six genomes under investigation [Table 2.1-Table 2.6, Figure 2.3, Figure 2.4, Figure 2.5, Figure 2.6]. This implies that exons lengths are limited to a shorter range than introns length. These results are consistent with previous observations for eukaryotic genomes (Dorit *et al.,* 1990; Long *et al.,* 1995; Sakharkar *et al.,* 2005). It is interesting to see that exon lengths peak at 101-150 bp for H. *sapiens, P. troglodytes, M.musculus, D. rerio, C. elegans* and *D. melanogaster* [Figure 2.3 and Figure 2.5]. These results suggest on a similar genome organizations in these genomes at the gene structure level. As expected such patterns are not observed for intergenic DNA suggesting for lesser constraints on these regions and their hyper-variability. These results further suggest that eukaryotic (vertebrate and invertebrate) genomes possess lower order physical organizations and have fine-scale architectures in addition to the higher order physical organizations like centromeres and distinct heterochromatic and euchromatic regions.

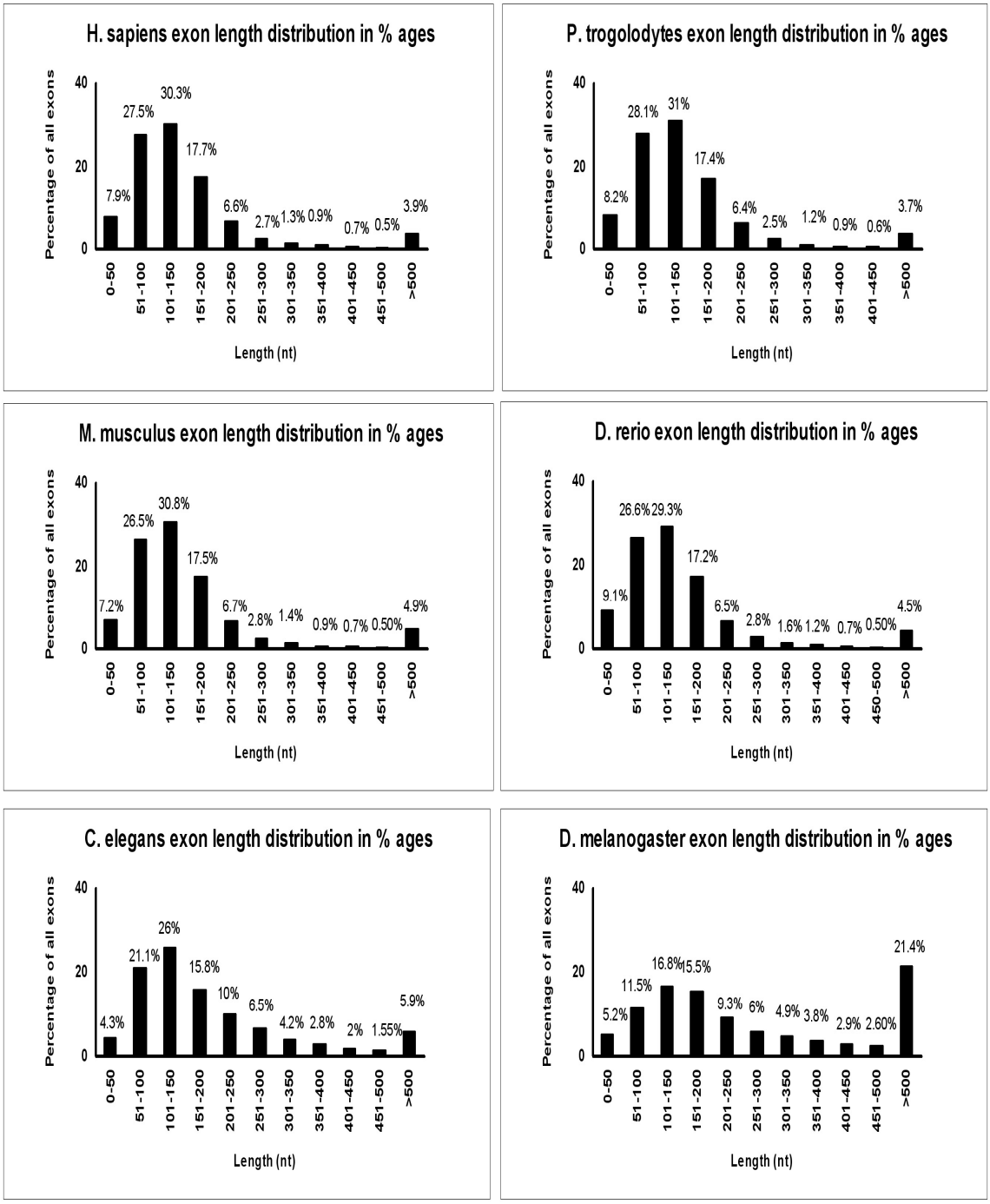

Figure 2.3
Exon length distribution in % for the six genomes

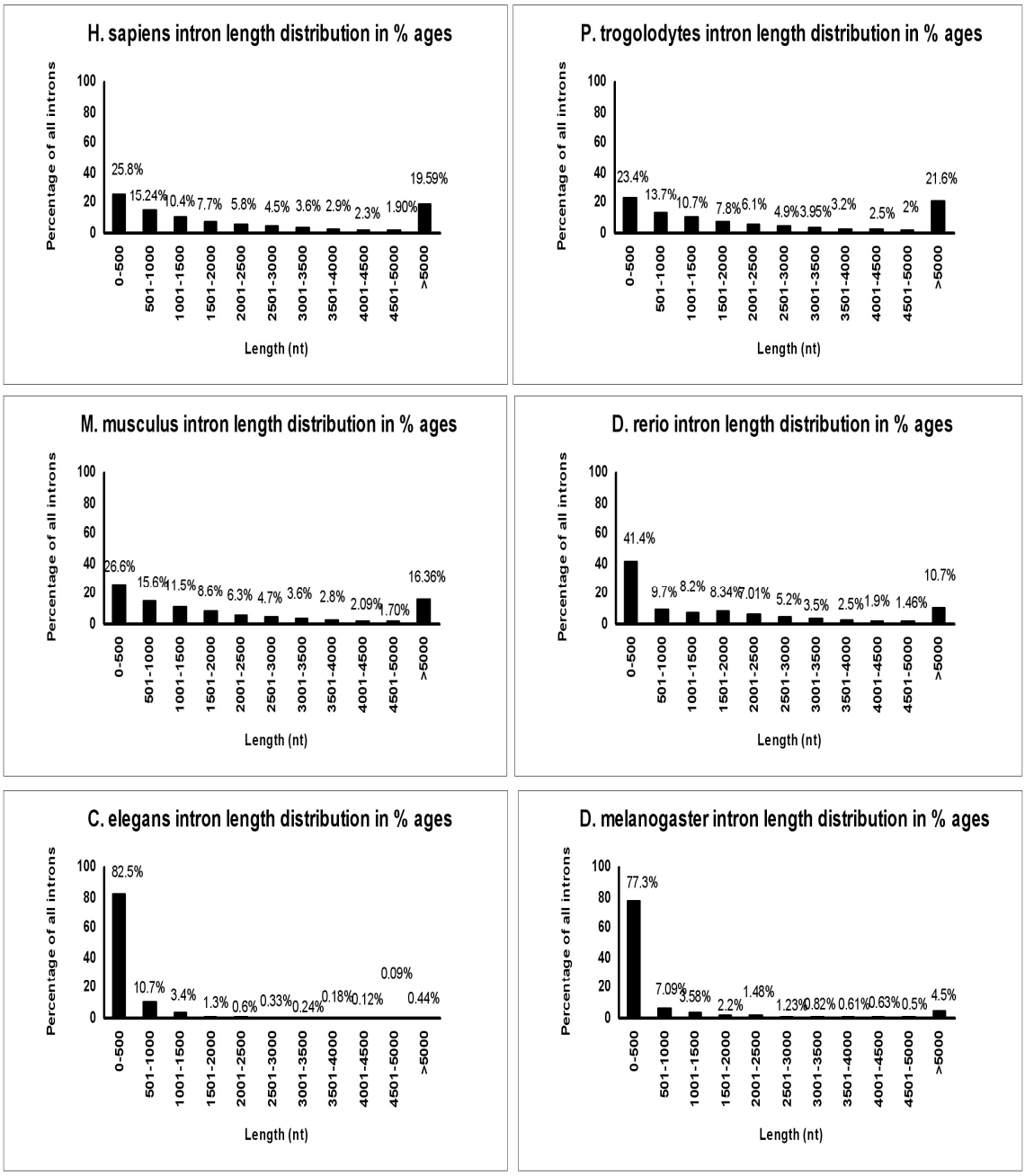

Figure 2.4
Intron length distribution in % for the six genomes

Average exon and intron length is 169 bp and 5231bp in *H. sapiens*; 158 bp and 5535 bp in *P. troglodytes*; 180 bp and 4529 bp in *M.musculus*; 175 bp and 2793 bp in *D.rerio*; 207 bp and 320 bp in *C. elegans* and 379 bp and 1022 bp in *D. melanogaster* genome. These data support the above observations on length constraints in exons and introns. These data are also supported by the fact that the standard deviations about the mean exon length are lesser than the standard deviations about the mean intron length. The greater standard deviations about the mean intron lengths suggests for their being under lesser selection pressures resulting in the tendency of large-scale changes which is reflected in their length distributions.

It is interesting to see that although, an intron can be thousands of base pairs in size, long introns make up only a small proportion of total introns in the genome [Table 2.7]. Less than 5.5% of introns are >20,000 bp in all the six organisms. Also, less than 1% of introns are <20bp in length in all the six genomes [Figure 2.4 and Figure 2.6]. These results suggest constraints on the splicing machinery to splice out very long or very short introns. More than 1/3rd of the genes in the two invertebrate genomes (*C. elegans* and *D. melanogaster*) have introns of length <500bp. These data suggest that *C. elegans* and *D. melanogaster*, the two invertebrates, do not contain long introns and <5% of introns in these genomes are >5000 bp (<1% for C. elegans). These data support earlier reports, suggesting that short introns in Drosophila melanogaster and Caenorhabditis elegans contain essentially all of the information for their recognition by the splicing machinery, and computer programs that simulate splicing specificity can predict the exact boundaries of ≈95% of short introns in both organisms.

Length in bp	% of introns in the organism					
	H.sapiens	*P.troglodytes*	*M.musculus*	*D.rerio*	*C.elegans*	*D.melanogaster*
0 - 500 bp	25.80%	23.42%	26.63%	41.39%	82.51%	78.65%
> 5000 bp	19.60%	21.61%	16.37%	10.72%	0.44%	0.12%
> 10000 bp	10.19%	11.01	8.14%	4.67%	0.08%	0.02%
> 15000 bp	6.69%	6.69%	5.39%	2.85%	0.02%	0.002%
> 20000 bp	4.97%	5.26%	3.95%	1.98%	0.005%	0.002%

Table 2.7
Intron length distributions along with their percentages

Earlier, it has been argued that, in vitro, for vertebrates, the assembly of ATP-dependent spliceosomes is inhibited if internal exons with strong constitutive splice sites are internally expanded to greater than 300 nucleotides (Robberson *et al.*, 1990). Concurrently, it has been proven that, In vivo, expansion of internal exons residing in vertebrate genes with moderate to large introns has two phenotypes: activation of internal cryptic splice sites within the expanded exon to create small exons or skipping of the entire exon (Berget, 1995). These phenotypes are consistent with splicing-imposed restriction on exon length. A few long vertebrate internal exons exist; the mechanism whereby such exons bypass restrictions on exon length is unknown. This suggests that a minimal separation between the splice sites might be required to prevent steric hindrance between the factors that recognize individual sites (Berget, 1995). Internal deletion of a constitutively recognized internal exon below 50 nucleotides is reported to result in skipping by the in vivo splicing machinery (Dominski and Kole, 1991). However, increasing the strength of the splice sites is reported to alleviate problems in recognition, suggesting that exon size and splice site strength are additive factors in exon recognition (Dominski and Kole, 1992). Very small exons have been reported to require special enhancing sequences in addition to strong splice sites for inclusion (*Black*, 1991; Black, 1992; Sterner and Berget, 1993). These enhancers are suggested to function as binding sites for splicing factors that

artificially extend the exon domain during exon recognition. Recently, a strong relationship between intron length and the prevalence of splice variation is reported.

The total intron length per kb of CDS averages to – 964.68, 968.99, 956.82, 932.97, 565.38, and 677.63 bp for *H. sapiens, P. troglodytes; M.musculus; D.rerio; C. elegans* and *D. melanogaster*, respectively. On the other hand, the numbers of introns per CDS are – 7.53, 8.33, 7.27, 6.92, 5.37, 3.52 for *H. sapiens, P. troglodytes; M.musculus; D.rerio; C. elegans* and *D. melanogaster*, respectively. These data suggest that *C. elegans* and *D. melanogaster* have contrasting intron-exon structures. *C. elegans* genes contain more (5.37 introns per CDS), shorter (565.38 bp each per kb of CDS) introns, while *D. melanogaster* genes have fewer, longer introns (3.52 introns per CDS, of length 677.63 bp per kb of CDS).

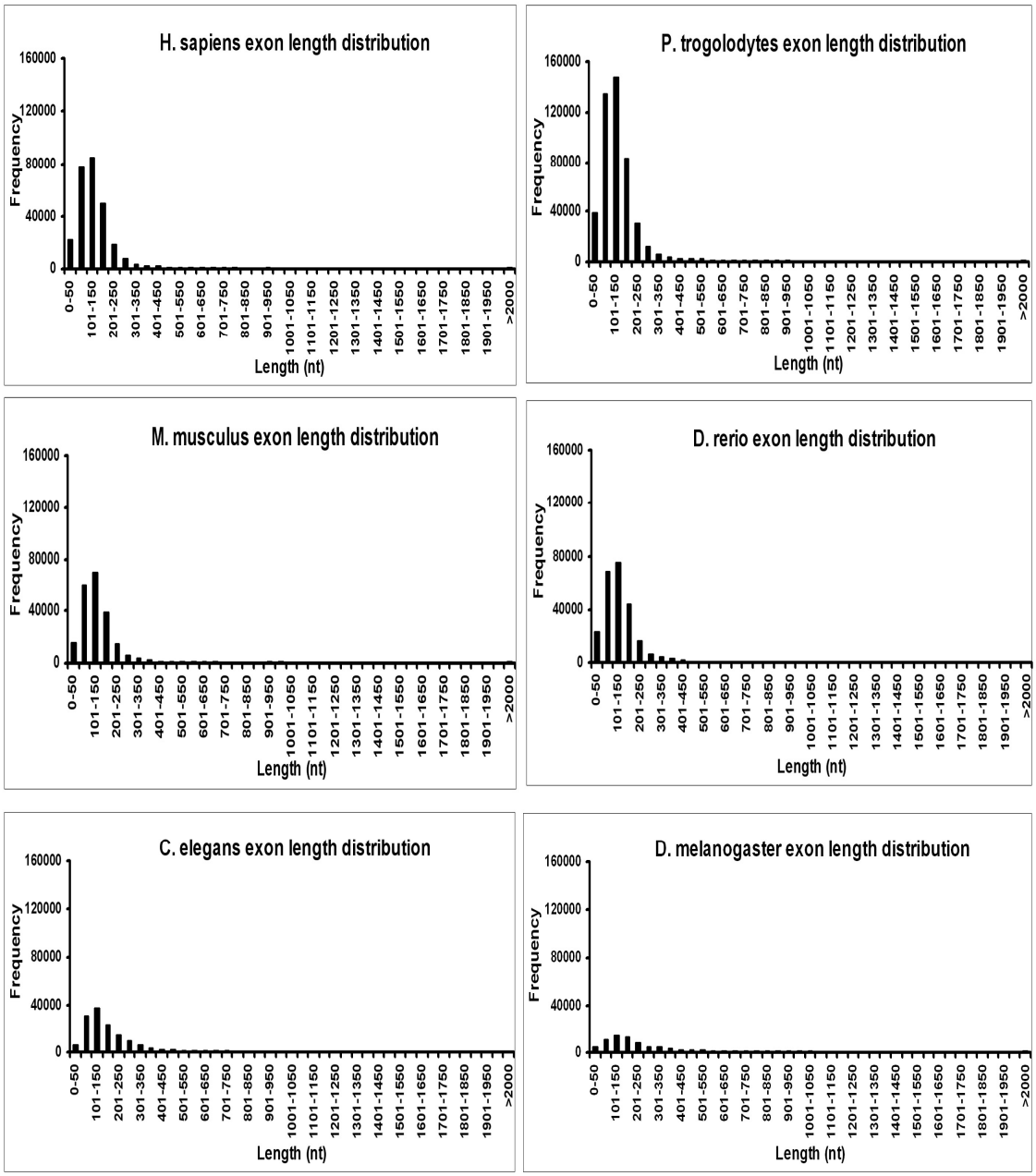

Figure 2.5
Exon length distribution in frequency for the six genomes

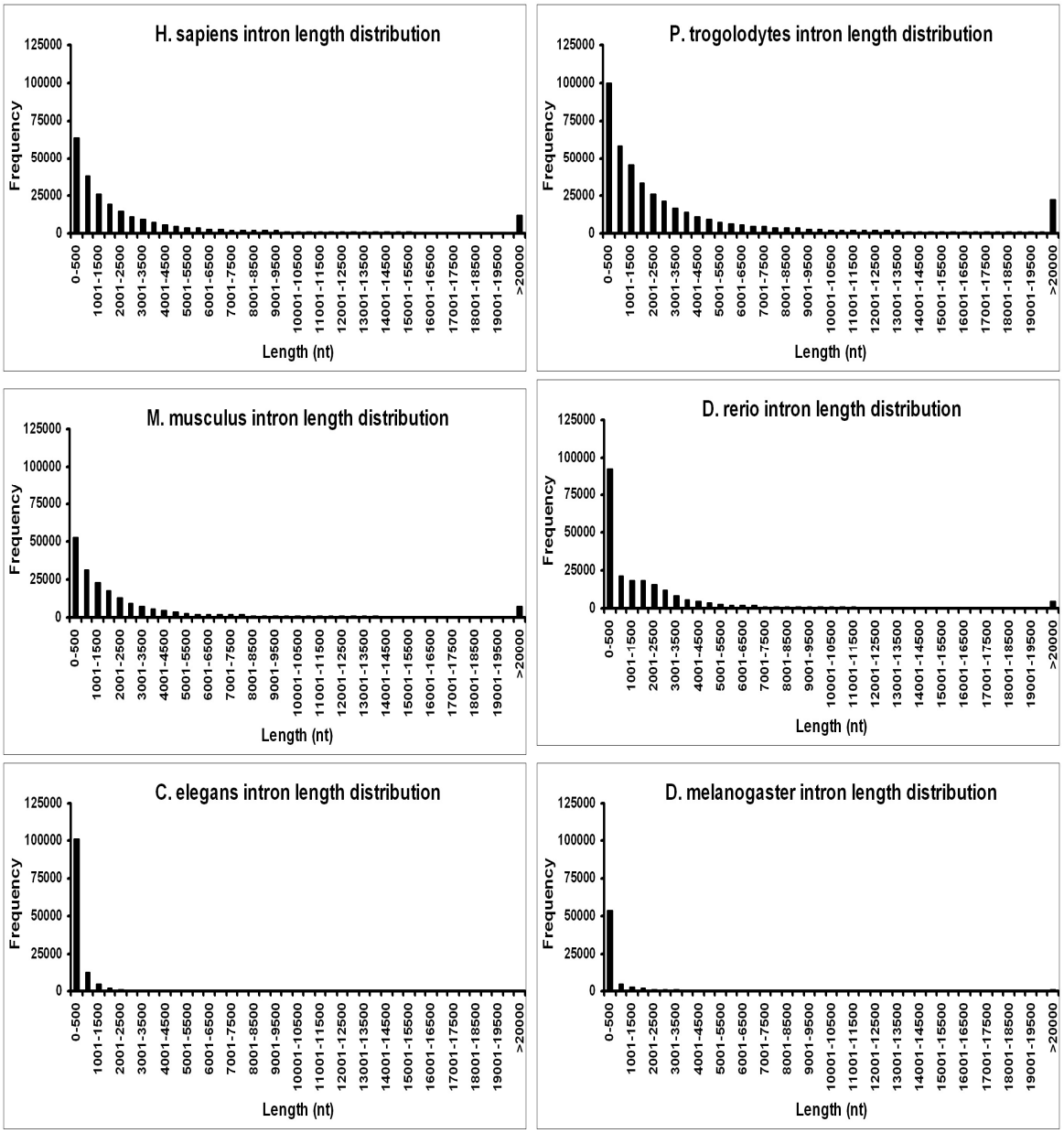

Figure 2.6
Intron length distribution in frequency for the six genomes

2.3.5 Correlations between Chromosome Size and Total Length in Exons, Introns and Intergenic DNA

It is observed that the correlation of total length in exons (bp) and chromosome size range from is 0.72, 0.74, 0.74, 0.87, 0.74, 0.99 for *H. sapiens*, *P. troglodytes; M.musculus; D.rerio; C. elegans* and *D. melanogaster*, respectively. The correlations between intron length and chromosome size are 0.96, 0.90, 0.89, 0.92, -0.16, 0.99 for these genomes. The correlation between intergenic DNA and chromosome size is 0.99, 0.97, 0.99, 0.96, 0.91, and 0.89 for these genomes. *C.elegans* also has the least number of intronless genes [Table 2.8].

Caenorhabditis elegans and Drosophila, normally only contain short introns. Average intron lengths in these organisms is 1000 and 600 nt, seperately. However, these two species have almost opposite intron-exon structures. Drosophila genes have fewer, longer introns (2.7 introns per kb of CDS, 564 nt each); on the other hand Caenorhabditis elegans genes have more (4.0 introns per kb of CDS), shorter (467 nt each) introns.

These results for suggest that in all the six genomes, for larger chromosomes more regions are covered in introns and intergenic DNA except in C. elegans. These observations indicate on the important role of introns and intergenic DNA in chromatin structure and chromosome architecture (since introns and intergenic DNA account for major component of the determined chromosome size (Venter *et al.*, 2001; Lander *et al.*, 2001). Lengyel and Penman showed that the size of hnRNA (heterogeneous nuclear RNA), but not mature mRNA, increases with genome size in

dipterans. This observation, dated before the discovery of the intervening sequences or introns in 1977, was the first indication of a positive relationship between genome size and total intron length (Lengyel and Penman, 1975). We also observed a positive correlation of, $r = 0.98$ for human, $r = 0.98$ for mouse, r=0.98 for chimpanzee and r=0.88 for zebra fish, 0.048 for *C.elegans* and 0.99 for *D. melanogaster* between total length in introns (bp) and total length in exons (bp).

Correlation	**Chr size**	**Exon**	**Intron**	**IGD**
H.sapiens				
Exon	0.72	X	X	X
Intron	0.96	0.81	X	X
IGD	0.99	X	X	X
P.troglodytes				
Exon	0.74	X	X	X
Intron	0.9	0.84	X	X
IGD	0.97	X	X	X
M.musculus				
Exon	0.74	X	X	X
Intron	0.89	0.81	X	X
IGD	0.99	X	X	X
D.rerio				
Exon	0.87	X	X	X
Intron	0.92	0.88	X	X
IGD	0.96	X	X	X
C.elegans				
Exon	0.75	X	X	X
Intron	-0.16	0.048	X	X
IGD	0.91	X	X	X
D.melangaster				
Exon	0.99	X	X	X
Intron	0.99	0.997	X	X
IGD	0.89	X	X	X

Table 2.8
Correlation values of intron, exon, igd size against chromosome size as well as correlation of intron size against exon size, for the six genomes.

These data hint that the proportions of the genome represented in exon and in intron are correlated in all the genomes (except *C. elegans*) and that perhaps the spatial requirement for regulatory DNA shaped the density of genes in these genomes. A

positive relationship between introns and genome size has now been established for many eukaryotes (Hughes and Hughes, 1995; Moriyama *et al*., 1998; Deutsch and Long, 1999; Vinogradov, 1999). In all cases, however, the differences in intron size alone cannot fully account for the differences in euchromatic genome size, indicating that a single class of non-coding DNA does not easily explain the differences in genome size. Our results imply that variation in genome size among organisms is usually associated to congruent changes across different classes of non-coding DNA (e. g. introns and intergenic regions) uniformly across the genome and propose a causal contribution by exons on each chromosome in genome design and architecture. These results confirm that not only the genome organization is similar in all the six genomes, but it is also non-random and is shaped by content in exons, introns and intergenic regions.

2.4 Summary

The result of the intron and exon length distribution analysis for all six eukaryotic genomes shows that the exon length are distributed more tightly than the intron length in general, which implies that, for eukaryotic genomes, coding regions are under more strict selection rules than non coding regions during the evolutionary process. The pattern of intron length distribution shows that the intron length has a larger scale change compared with the exon length but few introns are found to be very shot or very long, which means that even though intron is not under the selection pressure as high as exon, its length is still limited by the constraints of splicing machinery to splice out very long or very shot introns. The correlation study shows the proportions of the genome represented in exon and in intron are correlated in all the genomes (except *C. elegans*). The common characteristics of intron, exon length distribution

and correlation for the 6 eukaryotic genomes, provide meaningful understanding of eukaryotic genome architecture design, which helps for new eukaryotic genome structure prediction in future. Since this research is only conducted on six eukaryotic genomes, future similar studies can be performed with newer eukaryotic genomes being completely sequenced in NCBI, so that a deeper and broader insight of eukaryotic genomes evolutionary process can be achieved. With a larger number of eukaryotic genomes available, enough samples can be collected such that statistical and hypothesis tests can be conducted to prove what has been found.

CHAPTER 3 CHARACTERISTICS OF TARGETS WITH FDA APPROVED DRUGS

3.1 Background

Accumulated knowledge of genomic information, systems biology, and disease mechanisms provide an unprecedented opportunity to elucidate the genetic basis of diseases, and to discover new and novel therapeutic targets from the wealth of genomic data. With hundreds to a few thousand potential targets available in the human genome alone, target selection and validation has become a critical component of drug discovery process. The explorations on quantitative characteristics of the currently explored targets (those without any marketed drug) and successful targets (targeted by at least one marketed drug) could help discern simple rules for selecting a putative successful target. Here we use integrative in silico (computational) approaches to quantitatively analyze the characteristics of 133 targets with FDA approved drugs and 3120 human disease genes (therapeutic targets) not targeted by FDA approved drugs. This is the first attempt to comparatively analyze targets with FDA approved drugs and targets with no FDA approved drug or no drugs available for them. Our results show that proteins with 5 or fewer numbers of homologs outside their own family, proteins with single-exon gene architecture and proteins interacting with more than 3 partners are more likely to be targetable. These quantitative characteristics could serve as criteria to search for promising targetable disease genes.

The genomics revolution and advances in disease mechanisms and systems biology has provided a deluge of new potential targets for drug discovery (Loging *et al.,* 2007). Technological advances continue to be a central driving force in the

acceleration of the drug discovery process. High-throughput gene sequencing has revolutionized the process used to identify novel targets. Thousands of new gene sequences have been generated but only a limited number of these can be converted into validated targets likely to be involved in disease. The increased number of potential targets and the decreased amount of information is generating a bottleneck in the target validation process (Ofran *et al.,* 2006).

Several new and improved methods (Zheng *et al.,* 2006), and integrated and systems-based approaches (Lindsay, 2005; Sams-Dodd, 2005; Hardy and Peet, 2004), are being explored for identifying targets and druggable proteins. The commonly used computational methods have primarily been based on the detection of sequence and functional similarity to known targets [*Hopkins and Groom, 2002; Wang* et al., *2004*], drug-binding domain family affiliation [*Kramer, 2004; Wang* et al., *2004*], and structural analysis of geometric and energetic features (Hajduk *et al.,* 2004; Hajduk *et al.,* 2005). These methods are less effective in finding targets that exhibit no or low homology to known targets, disease proteins and proteins with available 3D structures. As such non-homologous and structurally unknown proteins constitute a substantial percentage, ~20–100%, of the open reading frames in many of the completed genomes and therefore, they are an untapped source of novel drug targets (Han *et al.,* 2004). Hence, methods independent of sequence and functional similarity, and structural availability, are highly desirable.

Han et al. described on the use of Support Vector Machine (SVM) algorithms and their potential applications for facilitating the discovery of innovative targets and reported that the prediction accuracy for non-druggable proteins is better than that of

druggable proteins. This probably results from the more diverse set of non-druggable proteins compared with that of druggable proteins, enabling SVMs to better recognize non-druggable proteins (Zheng *et al.,* 2006). Recently, Sakharkar et al. reported on the use of integrative analyses approaches and highlighted on the utility of large genomic databases for in silico (computational) systematic drug target identification in the post-genomic era (Sakharkar *et al.,* 2007). However, the two main bottlenecks in drug discovery and development are in identifying which protein targets may respond to drugs and which targets are relevant in disease. Also, there are a number of critical issues that must be considered as strategies are developed to elucidate the inherited determinants of targetability of a disease protein. In light of the above, there is a need to identify and quantify the characteristics of commercially available therapeutic targets, particularly with respect to those of the non-targeted disease proteins. Here, we quantitatively analyze the characteristics of 133 therapeutic targets (human disease genes) of FDA approved drugs and compare them with those of 3120 therapeutic targets that have no-FDA approved drugs or no drugs available for them. The possible common features of these targets are presented and discussed.

3.2 Method

The human disease genes list was downloaded from the GeneCards database (Safran *et al.,* 2002). GeneCards is an automated and integrated database of human genes, genomic maps, proteins, and diseases. 3253 genes were identified that are reported to be involved in human diseases (Dataset-1). We manually extracted the drugs available for the disease genes from the DrugBank database. DrugBank is a unique computational/cheminformatics resource that combines detailed drug (i.e. chemical) data with comprehensive drug target (i.e. protein) information. DrugBank combines the strengths of, PharmGKB, PubChem and Swiss-Prot to create a single, fully

searchable in silico drug resource that links sequence, structure and mechanistic data about drug molecules (including biotech drugs) with sequence, structure and mechanistic data about their drug targets (Wishart *et al.*, 2006).

Information on protein-protein interaction data was from Biogrid database (Stark *et al.*, 2006). This information could be derived for 1554 disease gene products; Tissues in which the genes are expressed from TissueDistributionDB (http://genome.dkfz-heidelberg.de/menu/tissue_db/index.html), level-4, tissue distribution data. A target is assumed to be primarily distributed in a tissue if no less than 8% of the total protein contents are distributed in that tissue. This information could be derived for 1924 therapetic targets; Pathways in which the disease gene products are involved is derived from the SwissProt knowledgebase (Boeckmann *et al.*, 2003). Pathway information could be derived for 1159 therapeutic targets. We further performed Protein family assignments using SwissProt to the proteins for the disease genes and their BLASTP homologs (at a cutoff value of 0.001) (Finn *et al.*, 2006). 2276 therapeutic targets (human disease gene products) could be assigned to a protein family. Further categorization of homologs was performed to identify the number of unique protein families a protein was homologous to. Gene architecture information pertaining to the number of exons was extracted from CCDS (Consensus CDS) database. This information could be extracted for 2087 disease genes. We further divided the 3253 human disease genes (therapeutic targets) into two datasets one with FDA approved drugs (FDA) and the other set with no FDA approved drugs or no drugs available for them (here on referred to as ‘no-FDA’). Statistical analyses were performed as described below to provide clues to the differences in the properties for disease genes with FDA approved drugs and disease genes with no-FDA approved

drugs [Table 3.2]. A test of determining the confidence interval for the difference of population proportions was performed on these two sets to check for differences at 95% confidence level, whereby: x is the number of proteins/genes with FDA approved drugs within the cut-off region (for example pathway <=1). m is the total number of genes (both FDA and no-FDA) within the cut-off region. y is the number of genes with FDA approved outside the cut-off region (for example pathway >1). n is the total number of genes (both FDA and no-FDA) outside the cut-off region. P1=x/m; and P2=y/n For 95% confidence level: Z=1.96; Then the sampling error E is given by The upper and lower limits of a 95% confidence interval for P1- P2 are between x/m - y/n -E and x/m - y/n +E

$$E=Z^* \sqrt{\frac{P_1 * (1-P_1)}{m} + \frac{P_2 * (1-P_2)}{n}}$$

Note: This formula is based on the following
assumptions:
1. The population proportions P1 and P2 are not too close to 0 or 1.
2. Two random samples are taken, one for each population, and the two samples are independent.
3. Sample sizes m and n are large.
The results of our analyses are presented.

3.3 Result

3.3.1 Mapping Drugs to Targets

The mapping of FDA approved drugs to genes involved in diseases (targets) identified 133 unique targets with 289 distinct (non-duplicate) FDA approved drugs [Table 3.1]. These results clearly suggest that one target may have multiple drugs that are reported as binding to it. An analyses on targets available for no-FDA approved drugs shows that, there are 385 disease genes targeted by 684 distinct no-FDA approved drugs

(investigational agents). It should be noted that based on data available in Drugbank 2735 disease genes are not targeted by any drug or investigational agent [Table 3.1].

Characteristic	Genes with drugs	# of Drugs
FDA approved	133	289
Not-FDA approved	385	684
No drugs	2735	0
Total	3253	

Table 3.1
Number of targets with and without drugs

3.3.2 Mapping Targets to Pathways, and Tissue Information

The mapping of targets with FDA approved drugs and targets with no-FDA approved drugs onto SwissProt knowledgebase and TissueDB was performed, to extract information on the number of pathways a target is involved in and the number of tissues a target is expressed. The distribution of pathway frequency for percentage of targets with FDA drugs and proteins with no-FDA drugs and number of tissues a target is expressed in, is shown in Figure 3.1 and Figure 3.2, respectively. Our results shows that, the targets with FDA approved drugs and targets with no-FDA approved drugs, when compared, show no significant bias in the number of pathways involved and the number of tissues a target is expressed (p value = 0.05, implies 95% confidence level) [Table 3.2].

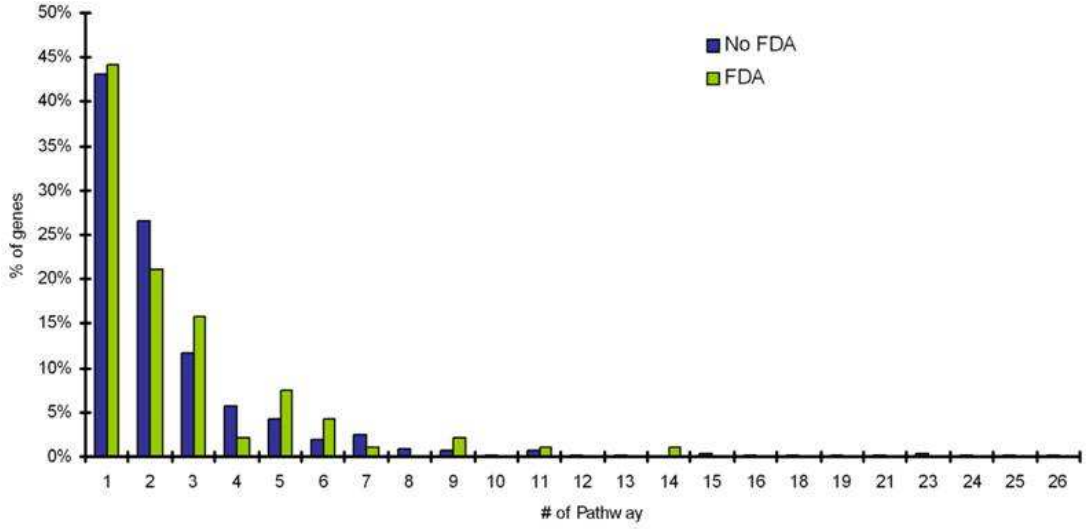

Figure 3.1
Distribution of pathway frequency for percentage of targets with FDA drugs and proteins with noFDA drugs. The number of pathways is shown along X axis and Y axis represents the % of genes involved in diseases (targets) with FDA approved drugs and noFDA approved drugs. It is interesting to see that more than 40% of targets are involved in only 1 pathway.

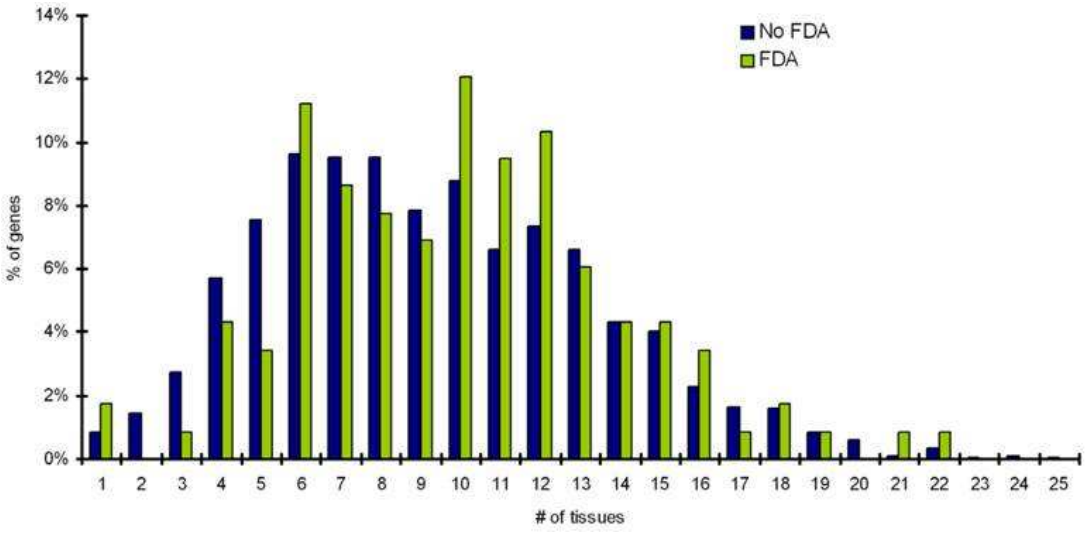

Figure 3.2
Distribution of number of tissues a target is expressed in for percentage of targets with FDA drugs and targets with noFDA approved drugs.

Type	Cutoff value	x	m	y	n	P_1	P_2	E	P_1-P_2-E		P_1-P_2+E	Result
Pathway	pathway =2	62	803	33	356	0.0772105	0.092697	0.035333	-0.0508192	<p1-p2<	0.0198469	no difference
Pathway	pathway =3	77	943	18	216	0.0816543	0.083333	0.040793	-0.0424721	<p1-p2<	0.039114	no difference
Pathway	pathway =4	79	1006	16	153	0.0785288	0.104575	0.051259	-0.0773052	<p1-p2<	0.0252125	no difference
Tissue	tissue =2	2	43	114	1881	0.0465116	0.060606	0.063862	-0.0779562	<p1-p2<	0.0497674	no difference
Tissue	tissue =3	3	93	113	1831	0.0322581	0.061715	0.037563	-0.0670202	<p1-p2<	0.0081066	no difference
Tissue	tissue =4	8	201	108	1723	0.039801	0.062681	0.02935	-0.0522302	<p1-p2<	0.0064694	no difference
PPI	ppi =3	38	759	61	795	0.0500659	0.07673	0.024146	-0.0508099	<p1-p2<	-0.0025175	Opposite
PPI	ppi =4	43	889	56	665	0.048369	0.084211	0.025385	-0.0612268	<p1-p2<	-0.0104563	Opposite
Pfam	pfam homologus =5	123	1869	10	407	0.0658106	0.02457	0.018777	0.02246343	<p1-p2<	0.0600177	Significant
Pfam	pfam homologus =6	124	1919	9	357	0.064617	0.02521	0.019633	0.01977433	<p1-p2<	0.0590395	Significant
Pfam	pfam homologus =7	124	1947	9	329	0.0636877	0.027356	0.020696	0.0156357	<p1-p2<	0.0570285	Significant
Exon	#exon =1	14	104	108	1983	0.1346154	0.054463	0.066354	0.01379829	<p1-p2<	0.1465066	Significant
Exon	#exon =2	19	218	103	1869	0.087156	0.05511	0.038846	-0.0068	<p1-p2<	0.0708926	no difference
Exon	#exon =3	26	349	96	1738	0.0744986	0.055236	0.029568	-0.0103058	<p1-p2<	0.0488311	no difference

Note For 95% confidence level $Z_{0.05}$=1.96

Table 3.2

Statistical analyses on significance of characteristics in targets with FDA and targets with noFDA drugs. This table shows that targets with single exonic gene architectures and more than 3 interacting partners are significantly more likely to have an FDA approved drug. Targets with >5 homologs outside their own protein family are significant less likely to have an FDA approved drug.

3.3.3 Mapping of Pathways to Homologs, Protein-Protein Interaction Data and Gene Architecture Information

Pfam assignments were performed for targets with FDA approved, and targets with no-FDA approved drugs. A distribution of homologs outside the target's protein family for percentage targets with FDA drugs and targets with no-FDA drugs is shown in Figure 3.3. We observe that ~60% of targets with FDA approved drugs have no homologs outside their own Pfam family. Our statistical calculation confirms that targets having ≤ 5 BLAST homologs outside its own Pfam family (at an e-value cutoff of 0.001) are more likely (p value=0.05) to be targetable (i.e. have FDA approved drugs available for them) than targets proteins having >5 BLAST homologs outside its own Pfam family [Table 1.2]. Mapping of targets to CCDS database shows that targets with single exon gene architectures are more likely to have FDA approved drugs available for them than targets with multi-exon gene architecture (p value =

0.05). Distribution for exon numbers for percentage of disease genes (targets) with FDA approved drugs and with no-FDA approved drugs is shown in Figure 3.4. Statistical analysis confirms that single exon genes are more likely to have FDA approved drugs (p value = 0.05). Statistical analyses on protein-protein interaction information from Biogrid database reveals that targets interacting with more than 3 partners are more likely to be targetable than proteins that do not interact with other proteins or interact with 3 or less number of partners [Table 3.2]. Distribution of Interacting partners for percentage of targets with FDA approved drugs and percentage of targets with no-FDA drugs is shown in Figure 3.5.

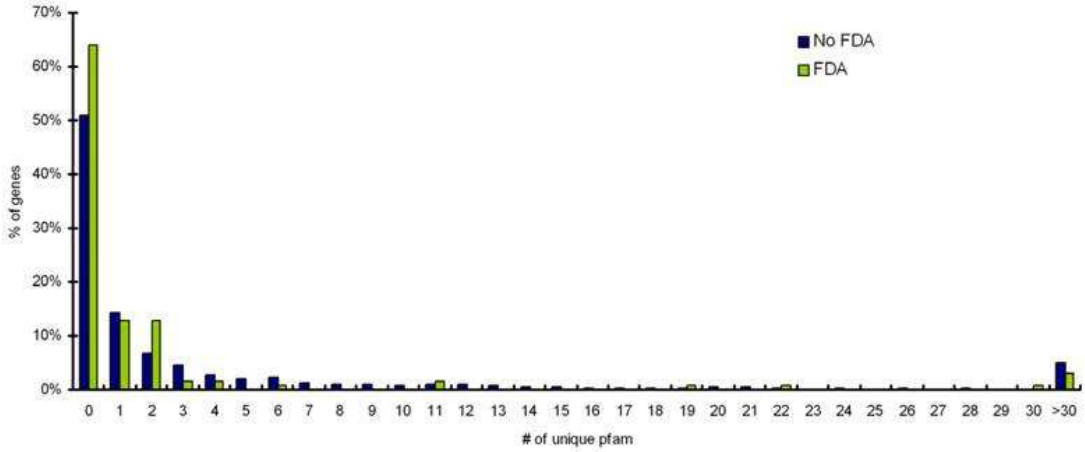

Figure 3.3
Distribution of homologs outside the target's protein family for percentage targets with FDA drugs and targets with noFDA drugs. Targets with less than 5 homlogs outside their own protein family are more likely to have FDA approved drugs available for them (p value =0.05).

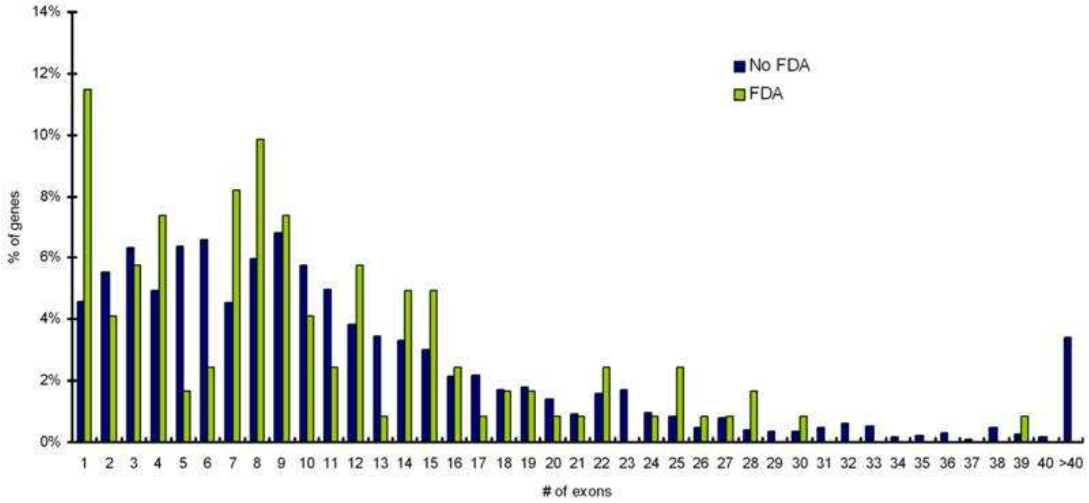

Figure 3.4
Distribution for exon numbers for percentage of disease genes (targets) with FDA approved drugs and with noFDA approved drugs. Targets with single exonic gene architectures are more likely to have FDA approved drugs available for them (p value = 0.05).

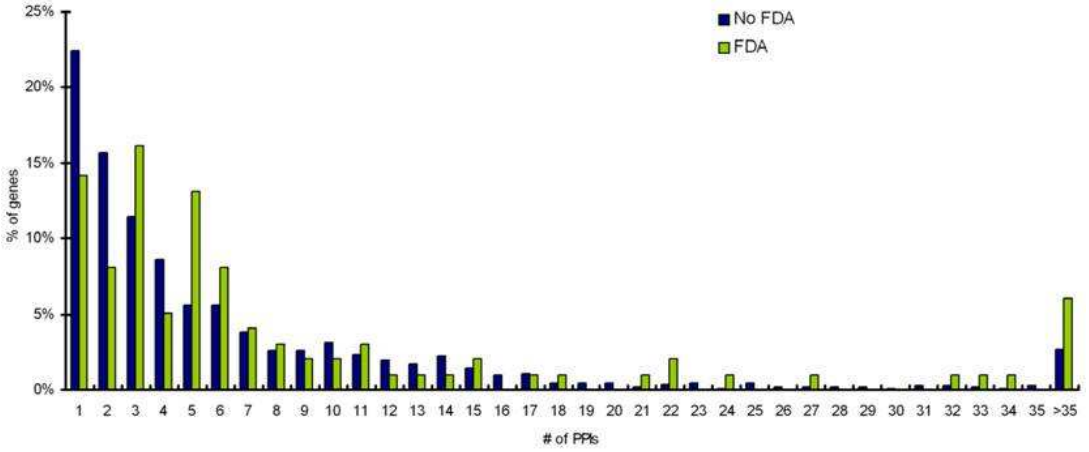

Figure 3.5
Distribution of Interacting partners for percentage of targets with FDA approved drugs and percentage of targets with noFDA drugs.

3.4 Discussion

It is well established that incorrect target selection accounts for the failures of some drug candidates. Experience from the biopharmaceutical industry indicates that currently only 5% of newly explored targets eventually lead to FDA-approved products (Caskey, 2007). Thus, innovative approaches to identify a “promising” target

are highly helpful in boosting productivity. Despite the clear need for better therapies for several disorders, novel drugs — particularly those that could revolutionize treatment — have been rare in recent years. Furthermore, biological complexity can often reveal unexpected and untoward effects of various treatment regimens (Mulder *et al.*, 1994; Nissen and Wolski, 2007). However, the pharmaceutical industry must provide innovative medications to treat disease. In the quest to find pharmacologic treatments for human diseases, many targets are screened, but in silico identification of efficacy remains essential towards shortening the path from target identification to verification of its efficacy as a target. Once a set of candidate proteins have been identified, suitability of a target for small-molecule or biological drug design is a key decision making criterion. The vast varieties of in silico resources that are available in life sciences research hold much promise towards aiding the drug discovery process. Here, we perform quantitative analyses to discern the salient features of the targets with FDA approved drugs. The principles learned from this exercise would serve as technical guidelines for better choices for the identification of putative successful targets leading to cost and time savings.

3.4.1 Pathway Affiliation

Collective actions of protein pathways are responsible for regulating disease processes and for response to drug actions. Therefore, the extent and specificity of target affiliation of pathways is likely to have statistically significant impact on drug actions (Zheng *et al.*, 2006). In both therapeutic targets with FDA approved drugs and therapeutic targets no-FDA categories the number of proteins decreases with an increase in the number of pathways [Figure 3.1]. This data suggests that involvement of a single protein in multiple pathways is not a preferred situation in a cell as this may lead to greater interference of a protein in related/non-related pathways and may

be a reason or cause for cross-reactivity. Comparison of therapeutic targets with FDA approved drugs and targets in no-FDA category (no FDA approved drugs or targets with no drugs available) suggests that the number of pathways a protein is involved in does not appear to be a factor determining its success as a target.

3.4.2 Number of Tissues

Although, highly selective tissue expression of a drug target, is attractive, as the potential for unwanted side effects may be more restricted, many effective drugs have been developed against targets that are widely expressed in the body (e.g. the angiotensin converting enzyme). Our analyses shows that as of today, based on targets with FDA approved drugs, the number of tissues a gene is expressed in does not affect the targetability of the protein [Table 3.2, Figure 3.2]. This is also complemented by the fact that localization of a gene in a particular tissue does not necessarily shed light on all the functions of that gene and at this stage in post-genomic era does not give sufficient details to infer any information in this direction. Since, in diseased tissue, gene expression levels often differ from those observed in normal tissues, with certain genes being over/under-expressed, or new genes being expressed or completely absent. Perhaps the most promising aspect is the information on differential expression in disease, since, the up or down-regulation of a gene may be the cause or result of the disease.

3.4.3 Protein Homologs outside Its Own Family

In the present day drug development processes, drug candidates have frequently been intentionally designed to bind to their target specifically and to avoid strong interactions with other human protein members of the same protein family to which the target belongs (Drews, 1997; Drews, 1997; Ohlstein et al., 2000; Terstappen and

Reggiani, 2001). However, their possible interactions with human proteins outside the family are not intentionally avoided at the design stage, and the potential unwanted effects associated with some of these interactions can only be detected at the later testing stages. Our analysis showed that proteins having more than 5 BLAST homologs outside its own family are significantly less targetable than proteins that have 5 or less number of homologs outside their own family (P=0.05) [Table 3.2, Figure 3.3]. This can be attributed to the fact that they can accommodate less target specific drugs that minimally interact with other pathways. Interactions with more number of targets may lead to secondary target effects, which may lead to cross-reactivity and unwanted interference. It is noteworthy that only 10 out of 133 targets with FDA approved drugs have more than 5 homologs outside their own protein family [Table 3.3]. Therefore, it tends to be easier to find successful drugs for those targets that have fewer human similarity proteins outside of their family.

Gene name	#Pathway	#PPI	#Tissue	#Homologs	#Exon	FDA drugs
ABL1	6	50	5	45	11	Imatinib
EGFR	11	95	9	40	28	Erlotinib***Gefitinib
PDGFRB	9	32	9	39	22	Imatinib
KIT	3	27	6	39	21	Imatinib
NPR1	3	3	18	30	22	Nitroglycerin
LDLR	0	9	6	22	18	Porfimer***Methyl aminolevulinate
F10	1	12	8	19	8	Enoxaparin***Heparin
PLG	2	33	14	11	19	Aminocaproic Acid
F2	3	24	0	11	14	Argatroban***Enoxaparin***Heparin
PCGR1A	1	10	9	6	6	Porfimer***Methyl aminolevulinate

Table 3.3
List of genes with maximum (Top10) homologs outside their protein family and their characteristics. PPI=number of protein-protein interactions or # of interacting partners. # pathway = number of pathways a target is involved, # tissue = number of tissues a target is expressed, # of exons = gene architecture information for the target. FDA approved drug shows the list of FDA approved drugs available for the target. Targets may have more than one FDA approved drug available against them. Drugs are separated by ***.

3.4.4 Exon Number

Data reveals an over-representation of single-exonic genes, among genes that have FDA approved drugs available for them (P=0.05). This can be explained based on the fact that single-exon genes do not undergo alternative splicing and hence can be used as drug targets with less caution (Sakharkar *et al.,* 2002). These results also corroborate with the fact that a major proportion of druggable genes have been reported as G-protein coupled receptors (GPCRs), and a major proportion of which (GPCRs) have been reported to be single exonic (Gentles and Karlin, 1999). These data support the fact that integration of data on gene annotation and gene architecture for genes involved in diseases has the potential to contribute to drug discovery and will be a step towards designing of safe, efficacious and promising drug targets. Accurate information on gene architecture and gene annotation allows us to at least be informed on the issue of splice variants. Besides, alternative splicing information is also useful at many stages of the drug discovery process including anti-sense mediated silencing and RNA interference (RNAi) for knock-down or down-regulation of specific genes products or designing of knock-out mice (Levanon and Sorek, 2003). It must be however, noted that many of the computationally derived annotations in the databases are either minimal or incorrect (apart from a carefully manually-curated database such as Swiss-Prot). Also, as annotation of genes is provided by multiple public resources, using different methods, it results in information that is similar but not always identical. However, the database used in this study, the CCDS database overcomes these issues as it is a collaborative effort to identify a core set of human protein coding regions that are consistently annotated and of high quality [Table 3.2, Figure 3.4].

3.4.5 Number of Interacting Proteins

It is becoming increasingly clear that genes and protein interactions in complex biological networks with local and global properties and perturbations of these networks contribute to the disease state. Understanding of interacting proteins is of importance in cell physiology and for developing novel treatments against disease. Small molecules that occlude crucial binding site(s) may be sufficient for modulating protein interactions that occur over large surface area and can thus act as drugs. It is known that these versatile protein-protein interactions are central to many key biological pathways and thus are attractive targets for drug discovery. Our data suggests that proteins interacting with more than 3 partners are preferred drug targets (P=0.05) [Table 3.2, Figure 3.5]. However, for the drug discovery process, it is important to determine the dynamics of interactions involving proteins having multiple interacting partners as well as identifying interaction surfaces for each partner. Moreover, research to discover small-molecule drugs that target protein-protein interactions is still at an early stage. The top 10 targets (with FDA approved drugs) based on number of interaction proteins are listed in Table 1.4. It is noteworthy that 5 out of these have more than 3 homologs outside their own family. These results hint on the fact that the above described characteristics do not work collectively/together to determine the success of a target.

Gene name	#Pathway	#PPI	#Tissue	#Homologs	#Exon	FDA approved drugs
EGFR	11	95	9	40	28	Erlotinib***Gefitinib
AR	0	81	12	2	8	Testosterone***Bicalutamide***Flutamide***Oxandrolone
ESR1	0	81	11	2	8	Fulvestrant***Raloxifene***Medroxyprogesterone***Progesterone***Estradiol***Ethinyl Estradiol***Estramustine***Tamoxifen***Conjugated Estrogens
NR3C1	1	71	4	2	8	Hydrocortisone***Methylprednisolone***Budesonide***Mometasone***Betamethasone***Loteprednol Etabonate***Amcinonide***Dexamethasone
BCL1	7	55	4	0	2	Paclitaxel***Docetaxel
ABL1	6	50	5	45	11	Imatinib
RARA	0	34	0	2	8	Tazarotene***Adapalene***Alitretinoin***Isotretinoin***Tretinoin***Acitretin
PLG	2	33	14	11	19	Aminocaproic Acid
PDGFRE	9	32	9	39	22	Imatinib
KIT	3	27	6	39	21	Imatinib

Table 3.4
List of genes with maximum (Top10) interaction partners and their characteristics. PPI=number of protein-protein interactions or # of interacting partners, # pathway = number of pathways a target is involved, # tissue = number of tissues a target is expressed, # of exons = gene architecture information for the target. FDA approved drug shows the list of FDA approved drugs available for the target. Targets may have more than one FDA approved drug available against them. Drugs are separated by ***.

3.5 Summary

We have to keep in mind that the drug discovery and development process is extremely difficult due to our poor understanding of biology of the disease and biology of the host (i.e., Homo sapiens). We are making steady progress, but there is still a long way to go. Knowledge on characteristics of targets could be helpful for predicting features and if possible deriving rules that guide new drug design and the search for new targets from genomic data. Our analyses hint that proteins with 5 or fewer homologs outside their own family, proteins with single-exon gene architectures and proteins with more than 3 interacting partners are promising targets. For targets with a higher number of similarity proteins outside their own family, or multiple

exons or interacting with less than 3 partners, it is still possible to find drugs. The characteristics defined above, merely make the tasks for finding successful drugs against these targets easier. As of today there is only 1 successful target NPR1 that is targeted by Nitroglycerin (interacts with 3 proteins, has 22 exons and has 30 homologs outside its own protein family) that does not satisfy all the three characteristics defined above. Moreover, in the FDA list there are 10 targets with more than 5 homologs outside their own protein family, 72 targets with 3 or lesser interacting partners and 108 targets with multi-exon gene architectures. These results suggest that the above quantitative characteristics selectively function in a combined, collective and differential mode and may help define and determine the targetability of a protein.

CHAPTER 4 DATABASE ON MISMATCHED INTRONS (MIDB)

4.1 What Are Sliding Intron Positions

Sliding intron positions or discordant intron positions are those that are either closely located in homologous or orthologous genes or an intron that is present in one gene but not in any of its homologues or orthologues. We have constructed a database on discordant intron positions in Human and Mouse. The data aims at systematically collecting information about discordant intron positions from genome data of these organisms and organising it into a form useful for understanding the genomics and dynamics of introns thereby helping understand the evolution of genes. The interface will allow examining of intron movements and will allow mapping of intron positions from homologous or orthologous proteins onto a single sequence. The data is of potential use for molecular biologists in general and for researchers who are interested in gene evolution and eukaryotic gene structure. Analysis of this data will allow us to identify putative cases of intron sliding, a process where an intron position shifts during evolution.

4.2 Material and Methodology

The completely sequenced genome/proteome data files (protein.fa) for human and mouse were downloaded from NCBI website (ftp://ftp.ncbi.nlm.nih.gov). CD-HIT was performed on the downloaded protein sequences to identify sequences that are 80% identical and the homologous or orthologous clusters of the protein sequence were generated. The corresponding genome data files in GenBank files were then used to extract and map the corresponding exon and intron positions on these aligned clusters of genes (homologues and orthologues). These clusters were further

categorized into six groups to identify any relation between protein length, intron position, number of introns. These data are of interest for evolutionary biologists. The methodology is shown in flowchart [Fig 4.1]

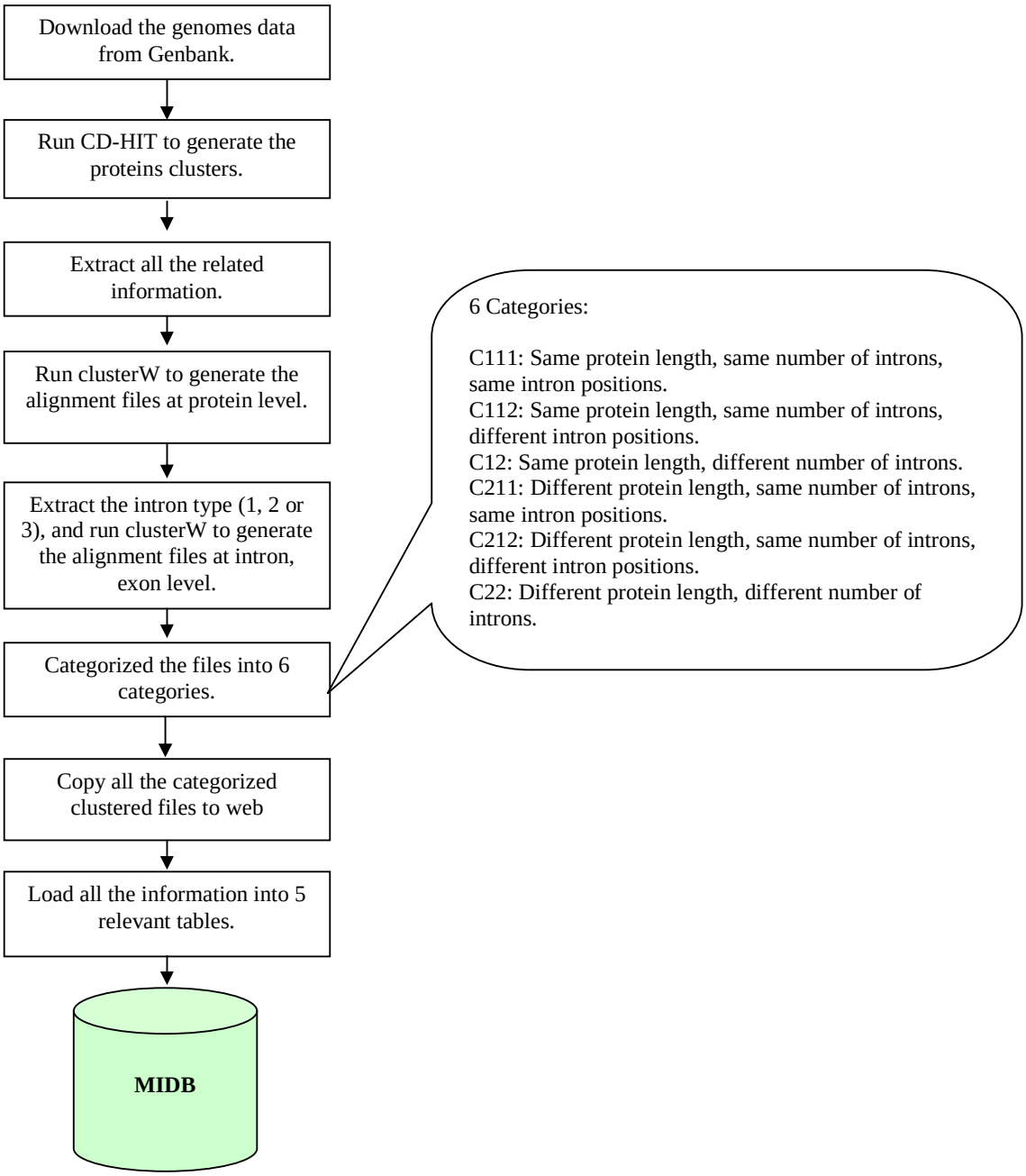

Figure 4.1
Methodology for creating MIDB.

The six categories are elaborated in detail below:

- C111: Same protein length, same number of introns, same intron positions.
- C112: Same protein length, same number of introns, different intron positions.
- C12: Same protein length, different number of introns.
- C211: Different protein length, same number of introns, same intron positions.
- C212: Different protein length, same number of introns, different intron positions.
- C22: Different protein length, different number of introns.

We further aligned the corresponding exons and intron for the clusters identified at 80% identity. Thus, we have protein alignments, exon alignments and intron alignments for these homologues or orthologues. These data were further stored in SQL database and made available through a web interface. The schema of the database is shown in Fig 4.2.

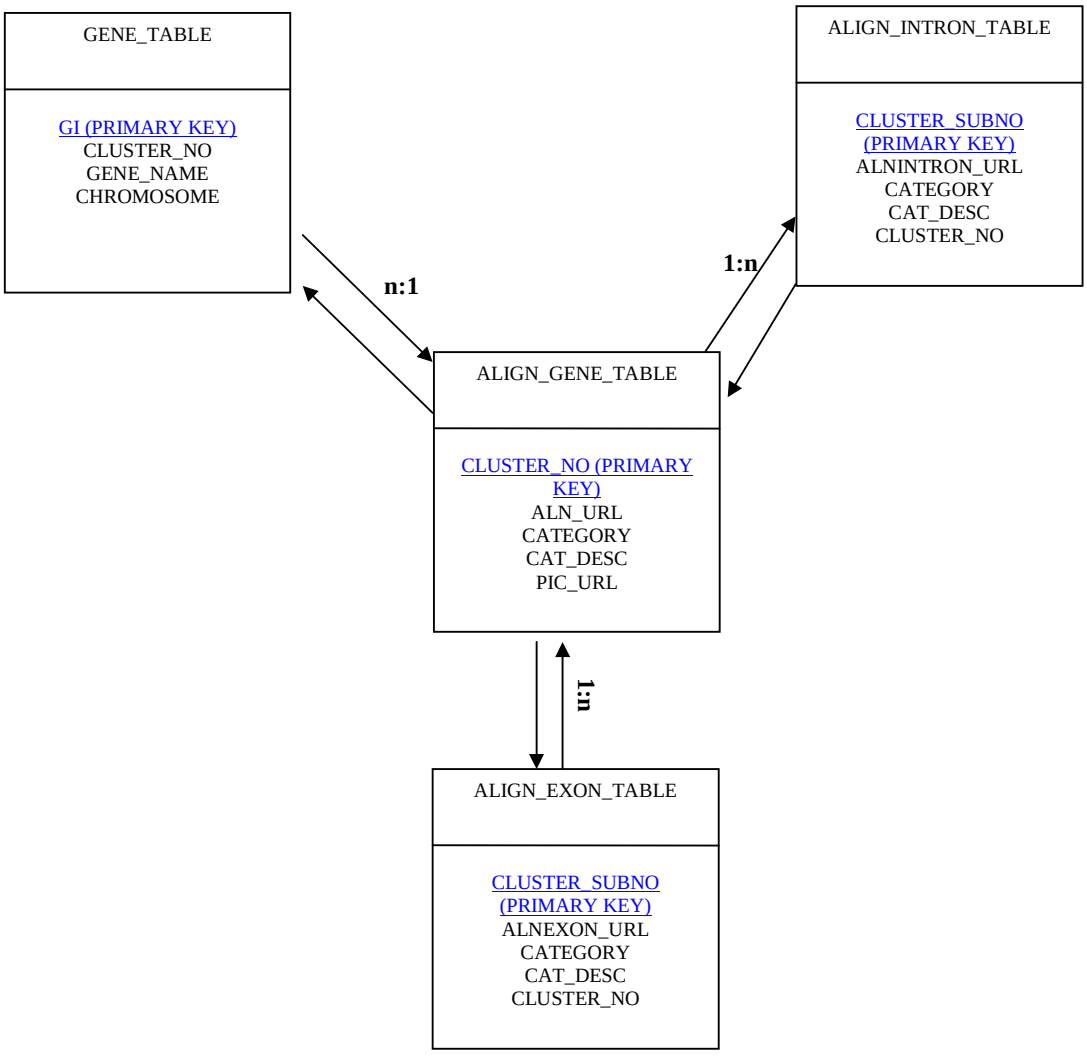

Figure 4.2
MIDB schema

Database Schema

The homologous and orthologous genes clusters and their intron and exon sequence mapping pictures are stored in the 4 tables shown in Figure 4.2.

GENE_TABLE: A look-up table for all the orthologous and homologous genes in human and mouse. The drop-down list of the search engine in the web is extracted from this table. It has four columns:

- GI: the unique identification number assigned to each gene in GeneBank.
- CLUSTER_NO: A unique number assigned to each orthologues or homologues.
- GENE_NAME: The name of gene.
- CHROMOSOME: The Chromosome number showing which it belongs to.

ALIGN_GENE_TABLE: The next level table of the gene_table, it stores all the information of the homologues and orthologues. The primary key is the CLUSTER_NO, and for each cluster, it records the web url pointing to the homologues and orthologues cluster sequence mapping. It is the core table, because it links the other 2 tables, align_exon_table and align_intron_table, which will be interpreted in the following paragraph. This table has 5 columns:

- CLUSTER_NO: The number assigned when generating the orthologue or homologue clusters using CD-HIT, for example, 13442.
- ALN_URL: The weblink of the cluster sequence alignment file on gene protein sequence level. It is named based on cluster number, such as 13442_aln.i.
- CATEGORY: Which category the orthologues or homologues belong to: C111, C112, C12, C211, C212 or C22?
- CAT_DESC: The description of the category.

- PIC_URL: A graphic view of the cluster alignment.

ALIGN_EXON_TABLE: Stores the exon information of the clustered orthologues or homologues. It links ALIGN_GENE_TABLE by CLUSTER_NO. Its primary key is CLUSTER_SUBNO, because for one cluster of genes, there can be multiple exons clusters. Two new columns are introduced in this table:

- CLUSTER_SUBNO: A unique number for each cluster of exons. For example, in cluster 13442, the orthologous genes all have 3 exons, and in this case it will have 3 clusters of exons in parallel with the protein cluster. They are 13442_e1, 13442_e2, 13442_e3.
- ALNEXON_URL: The weblink pointing to the clustered exon sequence mapping file.

ALIGN_INTRON_TABLE: Similar to ALIGN_EXON_TABLE, it stores the clustered intron information. Its CLUSTER_SUBNO records the numbers assigned to the clustered intron. Using the example above, they are 13442_i1, 13442_i2, because numbers of introns must be numbers of exons -1.

A User Guide to Search Engine

The significance of the MIDB is that it provides a platform for researchers who are conducting research on sliding introns, discordant introns or missing introns. Since it provides all the homologous and orthologous genes in human and mouse, by analyzing their sequence mapping at the protein, exon and intron level, users can simply identify the missing, sliding or discordant introns, so that the evolutionary rules of intron can be exposed. One simple example of how to use the search engine is illustrated below:

1. From the Search Engine drop-down list, select the gene of interest, for example, LOC646918. Clicking submit.
2. It will direct the user to the main interface with all the information, Figure 4.3. The first section of the webpage is a list of all the genes under the same cluster with LCO646918, and the data is at the gene level, including GI, chromosome number and gene name. The second section is the information at the protein level, which contains clustered protein sequence mapping, cluster category and graphic view of the mapping. The other two sections are for clustered intron and exon respectively, providing the link of intron and exon sequence mapping.
3. Users can proceed to click 14506 to view how the orthologues are aligned, as shown in Figure 4.4. The first 4 columns record the data of GI, chromosome number, protein length and gene name. One point to note is that |2| represents the location of the intron, and '2' stands for the phase of intron which has been discussed in chapter 1.
4. Similarly sequence mapping at the intron and exon level can be visualized by clicking 14506_i1, 14506_e1 or 14506_e2. A user who is studying introns can easily find out the changes or movement of the clustered introns. According to the intron sequence mapping data, most of the clustered introns have very large variance, even though they are 100% matched at the protein or exon level.

GI	CLUSTER	CHROMOSOME	GENE NAME
89036693	14506	human_ref_chrX	LOC646918
89037009	14506	mouse_ref_chrX	LOC651923
PROTEIN CLUSTER	CATEGORY		GRAPHIC VIEW
14506	C111: Same protein length, same number of introns, same intron positions.		GRAPHIC URL
INTRON CLUSTER			
14506_i1			
EXON CLUSTER			
14506_e1		14506_e2	

Figure 4.3
Demonstration on how to query MIDB.

89036693	human_ref_chrX	40	LOC646918	MDCRQVQCNS|2|SSNPIKDYIQE
89037009	mouse_ref_chrX	40	LOC651923	MDCRQVQCNS|2|SSNPIKDYTQE

Figure 4.4
Demonstration on protein level sequence mapping for one group of orthologues.

4.3 Data Analysis

Category	Number of aligned proteins in human & mouse	Number of 100 percent aligned proteins in human & mouse
C111	2697	77
C112	45	0
C12	111	6
C211	3248	10
C212	571	7
C22	4401	73
Total	11073	173

Table 4.1
Number of count of aligned proteins in Human and Mouse

The above table shows that about 11,000 proteins in human and mouse align at >80% amino acid sequence identity. 2,697 proteins have same protein length, same number of introns, same intron positions. 45 proteins have same protein length, same number of introns, different intron positions. 111 proteins have same protein length, different number of introns. 3,248 proteins have different protein length, same number of introns, same intron positions. 571 proteins have different protein length, same number of introns, different intron positions. 4.401 proteins have same protein length, different number of introns.

The striking result of this analysis was the number of genes for which all intron positions matched exactly. Of 11,073 orthologous human–mouse pairs, 5945 (C111, C211) showed no deviations at all in intron position alignment. These data suggest that comparative methods employing information about gene structures should be very successful in correctly predicting exon boundaries in genomic sequences. These results are also important in informing comparative gene prediction. If orthologs between human and mouse have virtually identical intron–exon structures, then cases of ambiguous assignment of intron boundaries should be resolvable by comparison with the other species. These results are an important step in the debate over the

relative roles of various processes in the shaping of the modern intron–exon structures of genes. Intron displacement or sliding is critically important for explaining the present distribution of introns among orthologous and homologous genes. MIDB allows examining of intron movements and allows mapping of intron positions from homologous or orthologous proteins onto a single sequence. The database is of potential use for molecular biologists in general and for researchers who are interested in gene evolution and eukaryotic gene structure.

CHAPTER 5 CONCLUSION AND RECOMMENDATIONS

5.1 Contribution

The explosion of data on all levels of the biological continuum made possible by the new developments in (a) technique, (b) instrumentation, (c) representation, storage and distribution by GenBank, EMBL, and DDBJ is a source of both exhilaration and anxiety. The phrase "genomic revolution" has become commonplace, and is certainly appropriate in a purely quantitative sense. The impending availability of complete genomic sequences for the model organisms has ignited a new era in biomedical research that provides an unprecedented view of the genetic essence of multi-cellular organisms. This has launched biology onto the path of becoming a data-bound science, a “science” in which all the data of a domain—such as a genome—are available before the laws of the domain are understood. A point in the history of biology has been reached where new generalizations and higher order biological laws are being approached but obscured by the simple mass of data and its related diversity and heterogeneity. Of primary concern to many biologists is how best to organize this massive outpouring of data in a way that would lead to deeper theoretical insight, perhaps even a unified theoretical perspective for biology. The proposed marriage between biology and information technology in a productive way is cardinal for knowledge discovery from information repositories. In conformity with this, it is important to connect to the information and learn how to sift through it for the parts we need. To cater for this, it is imperative to build subject-based knowledge sets which allow one to (a) retrieve biological information from multiple, independently-managed sources and (b) integrate the retrieved data into data sets and derive from these knowledge sets biological knowledge about the field of interest. This design can be used in the automatic, adaptive, organization of knowledge in databases on the

Internet, facilitating the rapid dissemination of relevant information and the discovery of new knowledge. This thesis is an attempt to in this direction and helps us to understand the architecture of eukaryotic genomes and how the gene structure changes across same genes for closely related species. It is a step towards organizing the basic biological data in an user-friendly form and facilitates easy analysis of this data.

The generated datasets on eukaryotic gene structure for 6 eukaryotic genomes show that the proportions of the genome represented in exon and in intron are correlated in all the genomes (except *C. elegans*) and that perhaps the spatial requirement for regulatory DNA shaped the density of genes in these genomes. In all cases, however, the differences in intron size alone cannot fully account for the differences in euchromatic genome size, indicating that a single class of non-coding DNA does not easily explain the differences in genome size. Our results for the first time imply that variation in genome size among organisms is usually associated to congruent changes across different classes of non-coding DNA (e. g. introns and intergenic regions) uniformly across the genome and propose a causal contribution by exons on each chromosome in genome design and architecture. This information is of significance in understanding how genomes evolve and facilitates differential genome analysis. These results confirm that not only the genome organization is similar in all the six genomes, but it is also non-random and is shaped by content in exons, introns and intergenic regions.

Our analyses on proteins with FDA approved drugs shows that proteins with 5 or fewer homologs outside their own family, proteins with single-exon gene architectures and proteins with more than 3 interacting partners are promising targets. These results

have high implications and use in prioritizing eukaryotic drug targets, which is the first and most important step for drug discovery.

We have constructed a database (MIDB) on aligned proteins in human and mouse, include the alignment of their introns and exons. This database provides us the platform to study the sliding intron posions and how they affect the protein sequence and function in general. These data may further our understanding on how sequence variability is generated and how introns contribute to this.

5.2 Recommendations for Future Work

The current research on intron, exon length distribution is carried out only on 6 eukaryotic genomes which are completed sequenced. Future study can be extended to more eukaryotic genomes with full information. A larger number of samples collections can provide the possibility for statistical and hypothetical tests on the regularities being found. The findings proved by the tests are definitely more convincing and significant than the ones obtained only by observation and inference.

Future research on characteristics of targets for FDA approved drugs can be conducted to introduce more FDA candidates, and there could be other characteristics at genome design level that are not discovered in current research but may affect the drug ability.

It would also be of interest to extend the usage of MIDB in the direction of understanding how intron sliding can lead to splice variants. These results have implications in understanding how new genes or similar genes with different functions evolve. These data could further be used to understand eukaryotic genome evolution and duplicate genes.

PUBLICATIONS

Li Peng, Meena K. Sakharkar, Zhong Zhaowei. Quantitative analysis on the characteristics of targets with FDA approved drugs, *International Journal of Biological Sciences*. 2008, 4: 15-22.

Li Peng, Meena K. Sakharkar. Genome architecture: number, size and length distributions of exons and introns in six crown eukaryotic genomes, *International Journal of Integrative Biology*, 2009, 5(2), 87-102.

BIBLIOGRAPHY

Bagavathi, S. and Malathi, R. (1996). Introns and protein revolution--an analysis of the exon/intron organisation of actin genes. FEBS Lett, 392, 63-65.

Baxevanis, A.D. (2000). The Molecular Biology Database Collection: an online compilation of relevant database resources. Nucl. Acids, Res 28: 1-7.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S. and Haussler, D. (2004). Ultraconserved elements in the human genome. Science, 304, 1321-1325.

Belfort, M. (1991). Self-splicing introns in prokaryotes: migrant fossils? Cell, 11, 9-11.

Belfort, M. (1993). An expanding universe of introns. Science, 262, 1009-1010.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. (2000). GenBank. Nucl. Acids, Res 28: 15-18.

Berget, S. M. (1995). Exon recognition in vertebrate splicing. J. Biol. Chem, 270, 2411-2414.

Bertsekas, D. (1995). Dynamic programming and optimal control. Athena Scientific, Belmont, MA.

Black, D. L. (1991). Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non-neuronal cells? Genes, Dev. 5: 389-402.

Black, D. L. (1992). Activation of c-src neuron-specific splicing by an unusual RNA element in vivo and in vitro. Cell, 69: 795-807.

Blake, C. (1978). Do genes-in-pieces imply proteins in pieces? Nature, 273, 267.

Blake, C. (1979). Exons encode protein functional units. Nature, 277, 598.

Blake, C. (1983). Exons--present from the beginning? Nature, 306: 535-537.

Boeckmann B, Bairoch A, Apweiler R, Blatter M, Estreicher A, Gasteiger E, Martin MJ et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids, Res. 31: 365–70.

Buchman, A.R. and Berg, P. (1988). Comparison of intron-dependent and intron-independent gene expression. Mol. Cell Biol., 8: 4395-4405.

Burge, C. and Karlin,S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol., 268: 78–94.

Burge, C.B. and Karlin, S. (1998). Finding the genes in genomic DNA. Curr. Opin. Struct. Biol., 8: 346-354

Burset, M. and Guigo, R (1996). Evaluation of gene structure prediction programs. Genomics, 34: 353-67.

Carvalho, A.B. and Clark, A.G. (1999). Intron size and natural selection. Nature, 401, 344.

Caskey TC. (2007). The Drug Development Crisis: Efficiency and Safety. Annual Review of Medicine, 58: 1-16.

Cavalier-Smith, T.(1985). Selfish DNA and the origin of introns. Nature, 315: 283-284.

Cavalier-Smith, T (1991). Intron-phylogeny: a new hypothesis. Trends Genet, 7: 145-148.

Cech, T.R. and Bass, B.L.(1986). Biological catalysis by RNA. Annu Rev Biochem, 55: 599-629.

Cech, T.R. (1985). Self-splicing RNA: implications for evolution. Int. Rev. Cytol, 93: 3-22.

Cech, T.R. (1986). The generality of self-splicing RNA: relationship to nuclear mRNA splicing, Cell, 44: 207-210.

Chung, S., and Perry, R.P. (1989). Importance of introns for expression of mouse ribosomal protein gene rpL32. Mol. Cell Biol., 9: 2075-2082.

Churchill, G.A. (1989). Stochastic models for heterogeneous DNA sequences. Bull. Mathem. Biol., 51: 79-94.

Copertino, D.W. and Hallick, R.B. (1993). Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. Trends Biochem, Sci. 18: 467-471.

Crick, F.H. (1968). The origin of the genetic code. J. Mol. Biol., 38: 367-379

Crick, F.H. (1979). Split genes and RNA splicing. Science, 204: 264-271.

Darnell, J.E. (1978). Implications of RNA-RNA splicing in evolution of eukaryotic cells. Science, 202: 1257-1260.

Darnell, J.E. and Doolittle, W.F.(1986). Speculations on the early course of evolution. Proc. Natl. Acad. Sci., 83: 1271-1275.

Davuluri, R. V., Grosse, I. and Zhang, M. Q. (2001). Computational identification of promoters and first exons in the human genome. Nat. Genet, 29: 412-417.

De Souza, S.J., Long, M., Schoenbach, L., Roy, S.W. and Gilbert, W. (1996). Intron positions correlate with module boundaries in ancient proteins. Proc. Natl. Acad. Sci., 93: 14632-14636.

De Souza S.J., Long, M., Gilbert, W. (1996). Introns and gene evolution. Genes Cells, 6: 493-505.

Deutsch, M. and Long, M. (1999). Intron-Exon structures of eukaryotic model organisms. Nucl. Acids, Res. 27: 3219-3228.

Dibb, N. J. and Newman, A. J. (1989). Evidence that introns arose at proto-splice sites. EMBO, J. 8: 2015-2021.

Dibb, N. J.(1991) Proto-splice Site Model of Intron Origin. J. theor. Biol., 151: 405-416.

Dominski, Z. and Kole, R. (1991). Selection of splice sites in pre-mRNAs with short internal exons. Mol. Cell. Biol., 11: 6075-6083.

Dominski, Z. and Kole, R. (1992). Cooperation of pre-mRNA sequence elements in splice site selection. Mol. Cell. Biol.,12: 2108-2114.

Doolittle, W.F. (1978). Genes in pieces: were they ever together? Nature 272 581-582.

Dorit, R. L., Schoenbach, L. and Gilbert, W. (1990). How big is the universe of exons? Science, 250: 1377-1382.

Drews J. (1997). Proceedings of the Roche Symposium "The Genetic Basis of Human Disease". In: Drews J, Ryser S, eds. Human Disease—From Genetic Causes to Biochemical Effects. Berlin: Blackwell, 1997: 5–9.

Drews J. (1997). Strategic choices facing the pharmaceutical industry: a case for innovation. Drug Discov Today, 2:72–8.

Dujon, B. (1989). Group I introns as mobile genetic elements: facts and mechanistic speculations. Gene, 82: 91-114.

Early, P.W., Davis, M.M., Kaback, D.B., Davidson, N., Hood, L. (1979). Immunoglobulin heavy chain gene organization in mice: analysis of a myeloma genomic clone containing variable and alpha constant regions. Proc. Natl. Acad Sci., U S A 76: 857-861.

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, et al. (2006). Pfam: clans, web tools and services. Nucleic Acids, Res. 34: D247-51.

Galas, D. J. (2001). Sequence interpretation. Making sense of the sequence. Science, 291: 1257-1260.

Gelfand, M.S. and Roytberg, M.A. (1993). BioSystems, 30: 173-182.

Gentle, A.J. and Karlin S., (1999). Why are human G-protein coupled receptors predominantly intronless? Trends in Genet, 15: 47.

Gilbert, W. (1978). Why genes in pieces? Nature, 271: 501.

Gilbert, W. (1986). The RNA world. Nature, 319: 618.

Gilbert, W. (1987). The exon theory of genes. Cold Spring Harb Symp Quant Biol., 52: 901-905.

Gilbert, W. and Glynias, M.(1993). On the ancient nature of introns. Gene, 135: 137-144.

Grivell, L.A.(1994) Intron mobility. Invasive introns. Curr. Biol., 4: 161-164.

Hajduk PJ, Huth JR, Fesik SW. (2005). Druggability indices for protein targets derived from NMRbased screening data. J. Med. Chem., 48: 2518–25.

Hajduk PJ, Huth JR, Tse C. (2005). Predicting protein druggability. Drug Discov. Today, 10: 1675-82.

Hamer, D.H. and Leder, P.(1979). Splicing and the formation of stable RNA. Cell, 18: 1299-1302.

Han LY, Cai CZ, Ji ZL, Cao ZW, Chen YZ. (2004). Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. Nucleic Acids, Res. 32: 6437-44.

Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbruck, S., Lehmann, G., Luft, F., Reich, J. and Bork P. (1999). Alternative splicing of human genes: more the rule than the exception? Trends Genet, 15: 389-390

Hardison, R. C., Oeltjen, J. and Miller, W. (1997). Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome, Res. 7: 959-966.

Hardy LW, Peet NP. (2004). The multiple orthogonal tools approach to define molecular causation in the validation of druggable targets. Drug Discov. Today, 9: 117–26.

Hawkins, J.D., (1988). A survey on intron and exon lengths. Nucleic Acids, Res. 16: 9893-9908.

Hopkins AL, Groom CR. (2002). The druggable genome. Nat. Rev. Drug Discov., 1: 727–30.

Huang, M.T.F. and Gorman, C.M. (1990). Intervening sequences increase efficiency of RNA 3' processing and accumulation of cytoplasmic RNA. Nucl. Acids, Res. 18: 937-947.

Hughes, A. L. and Hughes, M. K. (1995). Small genomes for better flyers. Nature, 377 - 391.

Jacquier, A. (1990). Self-splicing group II and nuclear pre-mRNA introns: how similar are they? Trends Biochem, Sci. 15: 351-354.

Janssen, J.C., Hall, M., Fox, N.C., Harvey, R.J., Beck, J., Dickinson, A., Campbell, T., Collinge, J., Lantos, P.L., Cipolotti, L., Stevens, J.M. and Rossor, M.N. (2000). Alzheimer's disease due to an intronic presenilin-1 (PSEN1 intron 4) mutation: A clinicopathological study. Brain, 123: 894-907.

Jeffreys, A.J. and Flavell, R.A. (1977). The rabbit beta-globin gene contains a large large insert in the coding sequence. Cell, 12: 1097-1108.

Kramer R, Cohen D. (2004). Functional genomics to new drug targets. Nat. Rev. Drug Discov., 3: 965–72.

Kondrashov, A. S. and Shabalina, S. A. (2002). Classification of common conserved sequences in mammalian intergenic regions. Hum. Mol. Genet, 1: 669-674.

Krishna R. Kalari 1,2,3, Melanie Casavant 4, Thomas B. Bair 1,2, Henry L. Keen 1,2,5, Josep M. Comeron 6, Thomas L. Casavant 1,2,3,7 and Todd E. Scheetz (2006). First exons and introns - a survey of GC content and gene structure in the human genome. In Silico Biology, 6: 0022.

Lambowitz, A.M. and Belfort, M. (1989). Infectious introns. Cell 56 323-326.

Lambowitz, A.M.(1993). Introns as mobile genetic elements. Annu. Rev. Biochem., 62: 587-622.

Lander, E. S., et al., International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature, 409: 860-921.

Leder, A., Miller, H.I., Hamer, D.H., Seidman, J.G., Norman, B., Sullivan, M., Leder, P. (1978). Comparison of cloned mouse alpha- and beta-globin genes: conservation of intervening sequence locations and extragenic homology. Proc. Natl. Acad. Sci., 75: 6187-6191.

Lengyel, J. and Penman, S. (1975). hnRNA size and processing as related to different DNA content in two dipterans: Drosophila and Aedes. Cell, 5: 281-290.

Levanon EY, Sorek R. (2003). The importance of alternative splicing in the drug discovery process. Targets, 2: 109-14.

Lindsay MA. (2005). Finding new drug targets in the 21st century. Drug Discov. Today, 10: 1683-87.

Loging W, Harland L, Williams-Jones B. (2007). High-throughput electronic biology: mining information for drug discovery. Nat Rev Drug Discov, 6: 220-30.

Logsdon Jr., J.M. (1991). The recent origins of spliceosomal introns revisited. Curr. Opin. Genet, Dev. 8: 637-648.

Long, M., de Souza, S. J., Rosenberg, C. and Gilbert, W. (1998). Relationship between "proto-splice sites" and intron phases: evidence from dicodon analysis. Proc. Natl. Acad. Sci., USA 95: 219-223.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. Nat. Genet, 30: 13-19.

Morey, C. and Avner, P. (2004). Employment opportunities for non-coding RNAs. FEBS Lett., 567: 27-34.

Morgan, T. H.(1917). The theory of the gene. The American naturalist, 51: 513-544.

Moriyama, E. N., Petrov, D. A. and Hartl, D. L. (1998). Genome size and intron size in Drosophila. Mol. Biol. Evol., 15: 770-773.

Morl, M. and Schmelzer, C. (1990). Integration of group II intron bl1 into a foreign RNA by reversal of the self-splicing reaction in vitro. Cell, 60: 629-636.

Mount, S. M., Burks, C., Hertz, G., Stormo, G. D., White, O. and Fields, C. (1992). Splicing signals in Drosophila: intron size, information content, and consensus sequences. Nucleic Acids, Res. 20: 4255-4262.

Mueller, M.W., Allmaier, M., Eskes, R. and Schweyen, R.J.(1993). Transposition of group II intron aI1 in yeast and invasion of mithochondrial genes at new locations. Nature,366: 174-176.

Mulder BJ, van der Doef RM, van der Wall EE, Tijssen JG, Piek JJ, van der Meer J, et al. (1994). Effect of various antithrombotic regimens (aspirin, aspirin plus dipyridamole, anticoagulants) on the functional status of patients and grafts one year after coronary artery bypass grafting. Eur Heart, J. 15:1129-34.

Nesic, D. And Maquat, L. (1994). Upstream introns influence the efficiency of final intron removal and RNA 3'-end formation. Genes, Dev. 8: 363-375.

Nilsen, T.W. (1996). A parallel spliceosome. Science, 273: 1813.

Nissen SE, Wolski K. (2007). Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med., 356: 2457-71.

Ofran Y, Punta M, Schneider R, Rost B. (2005). Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. Drug Discov. Today, 21: 1475- 82.

Ohlstein EH, Ruffolo RR Jr, and Elliott JD. (2000). Drug discovery in the next millennium. Annu Rev Pharmacol Toxicol., 40:177–91.

Okazaki, Y., et al., FANTOM Consortium, RIKEN Genome Exploration Research Group Phase I & II Team (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature, 420: 563-573.

Orgel, L.E. and Crick, F.H.C.(1980). Selfish DNA: the ultimate parasite. Nature, 284: 604-607.

Palmer, J.D. and Logsdon Jr., J.M.(1991). The recent origins of introns. Curr. Opin. Genet., Dev. 1: 470-477.

Pandey, N.B. , Chodchoy, N., Liu, T.J. and Marzluff, W.F. (1990). Introns in histone genes alter the distribution of 3' ends. Nucl. Acids, Res. 18: 3161-3170.

Patthy, L.(1987). Intron-dependent evolution: preferred types of exons and introns. FEBS Lett., 214: 1-7.

Patthy, L. (1991). Modular exchange principles in proteins. Curr. Opin. Struct. Biol., 1: 351-361.

Patthy, L. (1991). Exons - original building blocks of proteins? BioEssays 13 187-192.

Patthy, L.(1994). Exons and introns. Curr. Opin. Struct. Biol., 4: 383-392.

Patthy, L. (1995). Protein Evolution by Exon-Shuffling Molecular Biology Intelligence Unit, R.G. Landes Company/Springer, New York.

Patthy L. (1999). Genome evolution and the evolution of exon-shuffling--a review Gene, 238: 103-114

Pennacchio, L. A. (2003). Insights from human/mouse genome comparisons. Mamm. Genome, 14: 429-436.

Perlman, P.S. and Butow, R.A.(1989). Mobile introns and intron-encoded proteins. Science, 246: 1106-1109.

Robberson, B. L., Cote, G. J. and Berget, S. M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol., 10: 84-94.

Ryu, W.S. , Mertz, J.E. (1989). Simian virus 40 late transcripts lacking excisable intervening sequences are defective in both stability in the nucleus and transport to the cytoplasm. J. Virology, 63: 4386-4394.

Safran M, Solomon I, Shmueli O, Lapidot M, Shen-Orr S, Adato A, et al. (2002). GeneCards 2002: towards a complete, object-oriented, human gene compendium. Bioinformatics, 18: 1542-43.

Sakano, H., Rogers, J.H., Huppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R. Tonegawa, S. (1979) Domains and the hinge region of an immunoglobulin heavy chain are encoded in separate DNA segments. Nature, 277: 627-633.

Sakharkar MK, Kangueane P, Petrov DA, Kolaskar AS, Subbiah S. (2002). SEGE: A database on 'intron less/single exonic' genes from eukaryotes. Bioinformatics, 18: 1266-7.

Sakharkar, M., Passetti, F., de Souza, J. E., Long, M. and de Souza, S. J. (2002). ExInt: an Exon Intron Database. Nucleic Acids, Res. 30: 191-194.

Sakharkar, M. K., Chow, V. T. and Kangueane, P. (2004). Distributions of exons and introns in the human genome. In Silico Biol., 4, 0032.

Sakharkar, M. K., Perumal1, B. S., Sakharkar2, K. R., and Kangueane1, P. (2005). An analysis on gene architecture in human and mouse genomes. In Silico Biology, 5, 0032.

Sakharkar MK, Sakharkar KR, Pervaiz S. (2007). Druggability of human disease genes. Int J Biochem Cell Biol., 39:1156-64.

Saldanha, R., Mohr, G., Belfort, M. and Lambowitz, A.M.(1993). Group I and group II introns. FASEB, J. 7: 15-24.

Sams-Dodd F. (2005). Target-based drug discovery: is something wrong? Drug Discov. Today, 10: 139-47.

Sands, A. T. (2003). The master mammal. Nat. Biotechnol., 21: 31-32.

Schaffner, W., Kunz G, Daetwyler, H, Telford J., Smith H O., Birnsteil, M L., (1978). Genes and spacers of cloned sea urchin histone DNA analyzed by sequencing, Cell, 14: 655-671.

Sharp, P.A. (1985). On the origin of RNA splicing and introns. Cell, 42: 397-400.

Sharp, P.A.(1994). Split genes and RNA splicing. Cell, 77: 805-815.

Snyder, E.E. and Stormo, G.D. (1993). Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. Nucleic Acids, Res. 21: 607-613.

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. (2006). BioGRID: A General Repository for Interaction Datasets. Nucleic Acids, Res. 34: D535-39.

Sterner, D. A., Carlo, T. and Berget, S. M. (1996). Architectural limits on split genes. Proc. Natl. Acad. Sci., USA 93: 15081-15085.

Sterner, D. A. and Berget, S. M. (1993). In vivo recognition of a vertebrate mini-exon as an exon-intron-exon unit. Mol. Cell. Biol., 13: 2677-2687.

Stormo, G.D. (2000). Gene-finding approaches for eukaryotes. Genome, Res. 10: 394-397.

Tarn, W.Y. and Steitz, J.A. (1996). Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. Science, 273: 1824-1832.

Terstappen GC and Reggiani A. (2001). In silico research in drug discovery. Trends Pharmacol Sci., 22: 23–6.

Tilghman, S.M., Tiemeier, D.C., Seidman, J.G., Peterlin, B.M., Sullivan, M., Maizel, J.V. and Leder, P. (1978). Intervening sequence of DNA identified in the structural portion of a mouse beta-globin gene. Proc Natl Acad Sci., 75: 725-729.

Tomita, M., Shimizu, N. and Brutlag, D.L. (1996). Introns and reading frames: Correlation between splicing sites and their codon positions. Mol. Biol. Evol., 13: 1219-1223.

Tycowski, Kazimierz, T., Mei-Di S. and Steitz A.J. (1996). A mammalian gene with introns instead of exons generating stable RNA products. Nature, 379: 464-466.

Uberbacher, E.C. and Mural, R.J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc. Natl Acad. Sci., 88: 11261–11265.

Venter, J. C., et al.. (2001). The sequence of the human genome. Science, 291: 1304-1351.

Vinogradov, A. E. (1999). Intron-genome size relationship on a large evolutionary scale. J. Mol. Evol., 49: 376-384.

Wang S et al. (2004). Tools for target identification and validation. Curr. Opin. Chem. Biol., 8: 371-77.

Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. and Lawrence, C. E. (2000). Human-mouse genome comparisons to locate regulatory sites. Nat. Genet., 26: 225-228.

Waterston, R. H., et al., Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520-562.

Watson, J. D. and Crick, F. H. C. (1953). A structure for deoxy-ribose nucleic acids, Nature, 171: 737-738.

West, J.L. and Halas, N.J. (2000). Applications of nanotechnology to biotechnology commentary. Curr. Opin. Biotechnol., 11: 215-217.

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids, Res. 34: D668-72.

Woods, I.G., Wilson, C., Friedlander, B., Chang, P., Reyes, D.K., Nix, R., Kelly, P.D., Chu, F., Postlethwait, J.H., and Talbot, W.S. (2005). The zebrafish gene map defines ancestral vertebrate chromosomes. Genome, Res. 15(9):1307-1314.

Woods, I.G., Kelly, P.D., Chu, F., Ngo-Hazelett, P., Yan, Y.-L., Huang, H., Postlethwait, J.H., and Talbot, W.S. (2000). A comparative map of the zebrafish genome. Genome, Res. 10(12):1903-1914.

Wu, T.D. (1996). A Segment-based Dynamic Algorithm for Parsing Gene Structure. J. Comp. Biol., 2: 375-394.

Zheng C et al. (2006). Progress and problems in the exploration of therapeutic targets. Drug Discov. Today, 11: 412- 20.