DEVELOPING BIOINFORMATICS TOOLS FOR THE STUDY OF ALTERNATIVE SPLICING IN EUKARYOTIC GENES

LIM YUN PING

SCHOOL OF MECHANICAL & PRODUCTION ENGINEERING

2005

TOOLS FOR STUDY OF ALTERNATIVE SPLICING

LIM YUN PING

Developing Bioinformatics tools for the study of alternative splicing in eukaryotic genes

Lim Yun Ping

School of Mechanical & Production Engineering

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

2005

Acknowledgments

I would like to express my sincere thanks to Assistant Professor Meena K. Sakharkar and Assistant Professor Pandjassarame Kangueane for giving me the opportunity to work under their supervision and guidance in this project. I also thank Professor Liew Kim Meow, Director, Nanyang Centre for SuperComputing and Visualization (NCSV) for his valuable suggestions and all possible support required for this project.

I would also like to thank Mr. Stephen Wong and Mr. Lai Loong Foong for their help and advice on cluster computing. My appreciation goes to Ms Iti Chaturvedi for helping with the database web interface.

My heartfelt thanks go to Mandar Chitre and my family for their support during this period.

This project utilizes the human genome data published in the public domain maintained by NCBI, NIH.

Table of Contents

Summary

Alternative splicing (AS) is a process that produces more than one protein isoform per gene in eukaryotes. It accounts for huge protein diversity in eukaryotic cells. Consequently, comprehensive studies on AS is of critical importance in cell biology. The protein diversity rendered by AS is combinatorially large and a number of mathematical models have been developed. Here we describe a procedure to identify human genes exhibiting AS by exon skipping (a common method of splicing). In this approach, we exclusively use full length cDNA sequences for the identification of spliced variants. Thus, we identified 1229 human genes exhibiting AS. Subsequently, a web based relational database (ASHESDB) is constructed to store and search these genes. The unique feature in ASHESDB is the collection of human genes exhibiting AS based exclusively on full length cDNA. This approach significantly reduces false positives in the database introduced by procedures determined using EST data.

Availability: http://sege.ntu.edu.sg/wester/ashes/.

List of Figures

List of Tables

Abbreviations

AS	alternative splicing
BLAST	Basic local alignment search tool
BLAT	blast-like alignment tool
CDS	coding region
cDNA	complimentary DNA
CGI	common gateway interface
CHR_FRAG	chromosome fragment
DESC	description
DNA	deoxyribonucleic acid
ESE	exonic splicing enhancer
EST	expressed sequence tags
FTPD	fronto-temporal dementia and parkinsonism
GBK	genbank
ID	identification number
ISS	intronic splicing silencer
KH	K homology domain
MGC	mammalian gene collection
mRNA	messenger ribonucleic acid
NCBI	national centre for biotechnology information
NIH	national institute of health
PERL	practical extraction resource language
POS	position
RefSeq	reference sequences
RRM	RNA recognition motif
SQL	structured query language
SR	serine arginine rich proteins
UTR	untranslated region
URL	uniform resource locator

Chapter 1

Introduction

1.1 Objective and scope

The key objectives of this project are:

1. To develop a model to identify human genes exhibiting alternative splicing by exon skipping.

2. To construct an online relational database for alternatively spliced human genes.

1.2 Organization of this report

The report is organized into six chapters. The first chapter states the objectives and explains the mechanism of alternative splicing and its significance. The second chapter describes the currently available methods, databases, and their development. It also describes a method developed in this project for identifying spliced variants by exon skipping. The third chapter describes the spliced variants identified by this procedure. The last chapter summarizes the project, explains the limitation in the methods described and suggests further improvements that may be implemented in the future.

1.3 Background

1.3.1 Central dogma of molecular biology and mRNA splicing in eukaryotes

The Central Dogma of Molecular Biology is explained in four major steps:

1. The DNA replicates its information by a process known as replication.

2. The DNA codes for the precursor messenger RNA (Pre-mRNA) during transcription. As shown in Figure 1, the pre-mRNA which is made up of protein coding segments (exons) and non-coding segments (introns).

3. In eukaryotic cells, the mRNA is processed (essentially by splicing) and modified in the nucleus by capping at the 5’ end followed by the addition of a poly (A) tail at the 3’ end to increase stability. The mature mRNAs are then exported from the nucleus to the cytoplasm.

4. Messenger RNA carries coded information to ribosomes. The ribosomes "read" this information and use it for protein synthesis by a process called translation.

1.3.2 Eukaryotic gene structure

Vertebrate genes are typically split into numerous small exons interrupted by much larger introns [51]. Introns are more prevalent in higher eukaryotes than lower eukaryotes. The genome size seems to be correlated with total intron length per gene [11]. For example, yeast introns are shorter than invertebrate introns, which in turn are shorter than those of human.

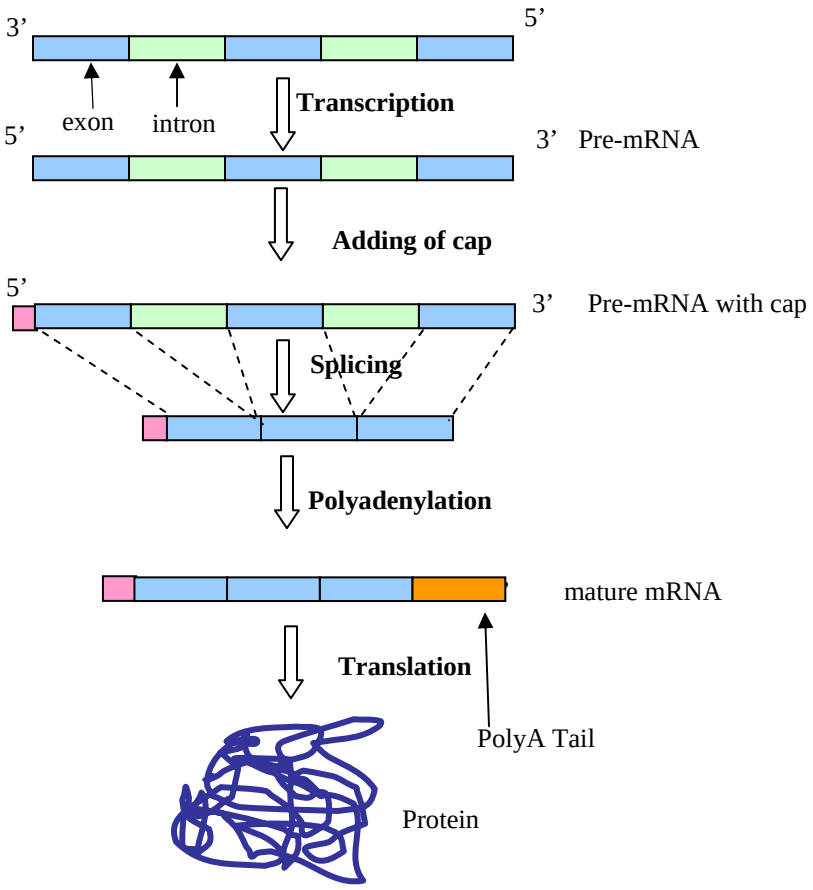

Figure 1. The central dogma of molecular biology.

During splicing, the introns are spliced from the pre-mRNA, leaving the exons to be translated into proteins.

A study revealed that the average length of human exons is 234 nucleotides long and introns is 1160 nucleotides long [39]. Introns smaller than 50 nucleotides, are significantly less frequent than longer introns, possibly due to a minimum intron size requirement for intron splicing. Among higher eukaryotes, exons have an extensive occurrence as they are found in most nuclear genes and often constitute more than 50% of pre-mRNA. In humans, exons only constitute 3 % of the genome [41].

1.3.3 The splicing reaction, machinery and spliceosome assembly

The primary transcript is processed by the spliceosome as shown in Figure 2, during which the spliceosomal introns are spliced out and the exons (coding segments) are spliced together to produce the mature mRNA. The spliceosomal introns that remained in the nucleus are rapidly degraded.

The spliceosome is a multi-component ribonucleoprotein complex that contains five small nuclear RNAs (snRNAs) and a large number of snRNA associated proteins.

Most introns start with GT or GU at the 5’ splice site and AG at the 3’ splice site. This represents the GT-AG rule of introns which holds in most, with exceptions being GC-AG and AT-AC. To date, two types of metazoan spliceosomes have been identified – namely an abundant U2-type (named after the U2 snRNA that base pairs with the donor site) and a minor U12-type (named after the U12 snRNA equivalent of U2 snRNA).

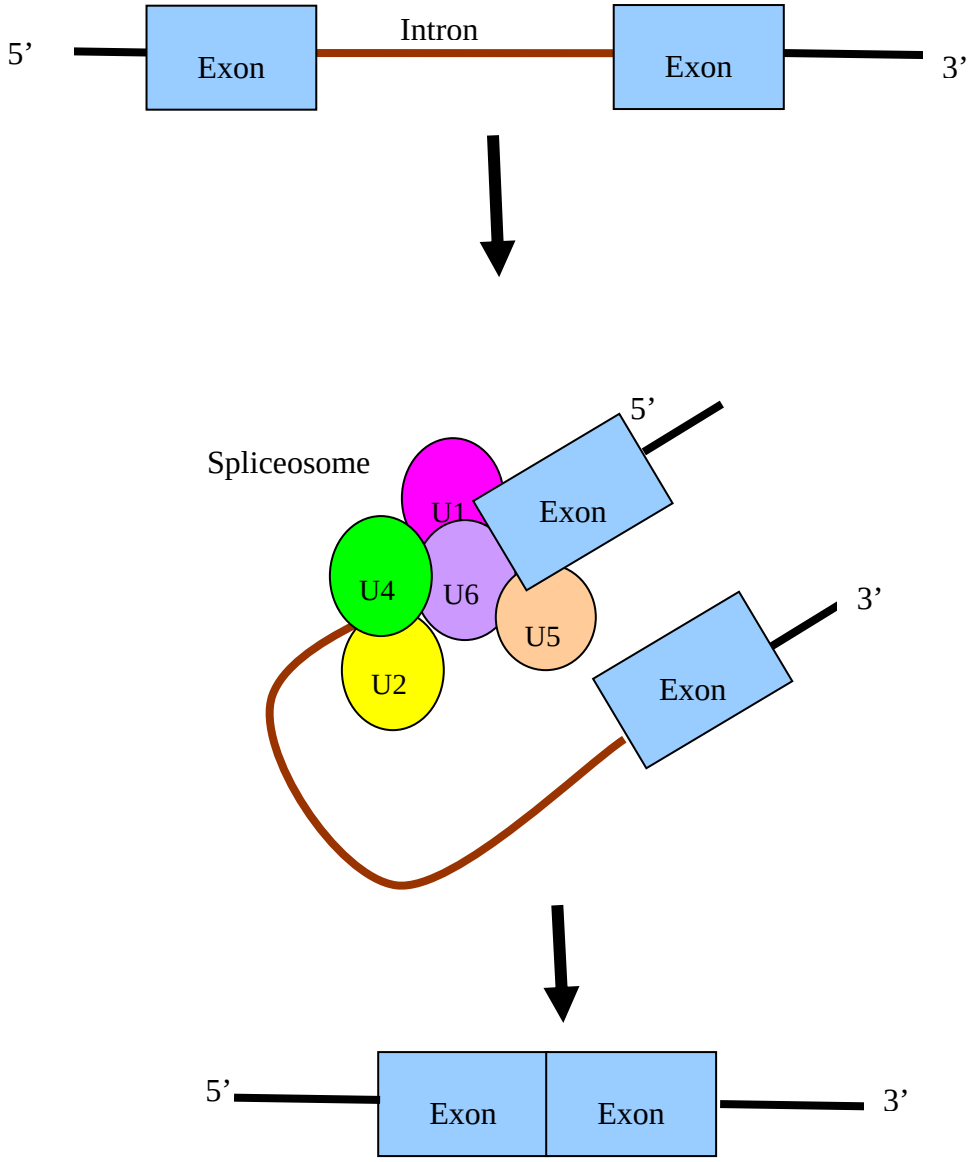

Figure 2. An illustration of the two- step splicing mechanism.

In the first step, the 5' end of the intron is joined to an adenine residue in the branch point sequence upstream from the 3' splice site to form a branched intermediate called an intron lariat. In the second step, the exons are ligated and the intron lariat is released.

Based on the consensus sequences and the type of spliceosome involved, introns can be sorted into two different groups. While the U2-type processes the GT-AG (and GC-AG) introns, the U12-type processes the minor AT-AC and the major GT-AG introns. [4,12].

Spliceosomes recognize 5' and 3' splice sites, which are located at exon-intron boundaries. The vast majority of pre-mRNA introns possess invariant GU (or GC) and AG dinucleotides at their 5’ and 3’ termini and are removed by the major spliceosome while the minor class introns with the non-canonical terminal nucleotides AU and AC at their 5’ and 3’ splice sites are recognized by the minor spliceosome [4]. The spliceosome is assembled upon the pre-mRNA sequence in a step-wise manner through interactions of its RNA and protein components with specific recognition sequences located on the pre-mRNA at the donor splice site, branch point, and the acceptor splice site (which includes a polypyrimidine tract and the terminal AG dinucleotide).

The 3’ splice site region consists of three sequence elements: the branch site, the polypyrimidine tract, and the terminal AG dinucleotides at the 3’ splice site boundary (AC dinucleotides in the case of minor class introns). These elements together provide a tight control for the 3’ splice site recognition. At the same time the diversity in the sequence composition of the branch site and the polypyrimidine tract in different pre-mRNAs provide abundant resources for the regulation of 3’ splice site selection in alternative splicing.

Alternative splicing is signaled by various weak signals (including regulatory sequences). The fidelity of splicing is ensured by the conserved sequence consensus. The splicing process in humans is also regulated by specific pentamer sequences within the intron

[37]. Purine and pyrimidine rich conserved sequence motifs, which are associated with exon skipping, have been identified using computational analysis [34].

Pre-mRNA splicing occurs in the spliceosome, a macromolecular complex which consists of five small nuclear ribonucleoprotein particles (snRNPs), designated as U1, U2, U4/U6 and U5 snRNPs, and a large number of non snRNP protein splicing factors.

The first step in spliceosome assembly is the formation of complex E, which is initiated by binding of the U1 snRNP to the 5'splice site and interactions of the SF1 and U2AF with the 3' splice site. Members of the SR family of splicing proteins can bind directly to the pre-mRNA and interact with the U1 protein and U2AF, thus bridging the splice sites. The formation of pre-splicing complex A involves the binding of U2snRNP to the branch site at the 3' splice site region and requires SF3a and SF3b to interact with the U2 snRNP. Then, U4, U6 and U5 tri-snRNP interacts with complex A to generate complex B, which is then converted into the active spliceosome.

Splicing occurs via a two-step mechanism. In the first step, the 5' end of the intron is joined to an adenine residue in the branch point sequence upstream from the 3' splice site to form a branched intermediate called an intron lariat. In the second step, the exons are ligated and the intron lariat is released.

The splicing of nuclear pre-mRNAs represents an essential step in the expression of genetic information [38]. The reaction requires the recognition of the exon-intron junctions and a precise juxtaposition of the splice sites within a catalytically active complex (the spliceosome) prior to phosphate bond cleavage. The spliceosome must be flexible enough to accommodate variations in intron length and a high degree of

specificity is necessary to coordinate the correct alignment of the cognate 5’ and 3’ splice sites. Another level of complexity is added in the case of alternative splice site selection which is often subject to tissue specific or developmental controls and can result in an on/off switch of gene regulation or in the generation of structurally or functionally distinct proteins from a single primary transcript by AS [43].

A fundamental problem in pre-mRNA splicing is ‘exon recognition’, the process by which exons are distinguished from introns, and intron-exon boundaries are precisely detected. The splicing machinery must recognize small exon sequences located within vast stretches of intronic RNA. Moreover, 5' and 3' splice sites are poorly conserved, and introns contain large numbers of cryptic splice sites, which match the loose 5' or 3' splice-site consensus. Cryptic splice sites are normally avoided by the splicing machinery, but can be selected for splicing when normal splice sites are altered by mutation. Identification of the correct splice sites is achieved by virtue of their proximity to exons. Specific sequence elements in exons known as exonic splicing enhancers (ESEs) interact with SR proteins, a family of conserved serine/arginine-rich splicing factors. These recruit the splicing machinery to the flanking 5’ and 3’ splice sites. Thus, exon sequences are under multiple evolutionary constraints, as they must be conserved not only for protein coding but also for recognition by SR proteins.

Once exon recognition is completed, the flanking splice sites must be joined in the correct 5' - 3' order to prevent exon skipping. This is accomplished, partly through the mechanistic coupling of transcription and splicing. Coupling transcription to splicing prevents exon skipping. Exon skipping occurrences in constitutively spliced pre-mRNAs,

are determined by the presence or absence of regulatory proteins that determine whether an exon is recognized and subsequently included in the mature mRNA. Mutations that interfere with proper exon recognition result in a large number of human genetic diseases [6]. Approximately 15% of the single base-pair mutations that cause human genetic diseases result in pre-mRNA splicing defect. Some of these mutations interfere with the function of normal 5' and 3' splice sites, thereby leading to the recognition of nearby pre-existing cryptic splice sites or creating new ones which are being used instead of the normal ones. Finally, single base-pair mutations within exons as shown in Figure 3, can interfere with the binding of SR proteins, leading to exon exclusion from the mature mRNA. For example, a translationally silent C-to-T mutation that occurs within an ESE of the human survival motor neuron 2 (SMN2) gene disrupts the binding site of the SR protein SF2/ASF and leads to exon skipping [6].

Regulatory proteins interact with specific sequences within pre-mRNAs and subsequently stimulate or repress exon recognition [33]. These proteins bind directly to 5' or 3' splice sites, or to other pre-mRNA sequences called exonic or intronic splicing enhancers (ESEs or ISEs) and silencers (ESSs or ISSs). Enhancers and silencers stimulate or repress splice-site selection, respectively.

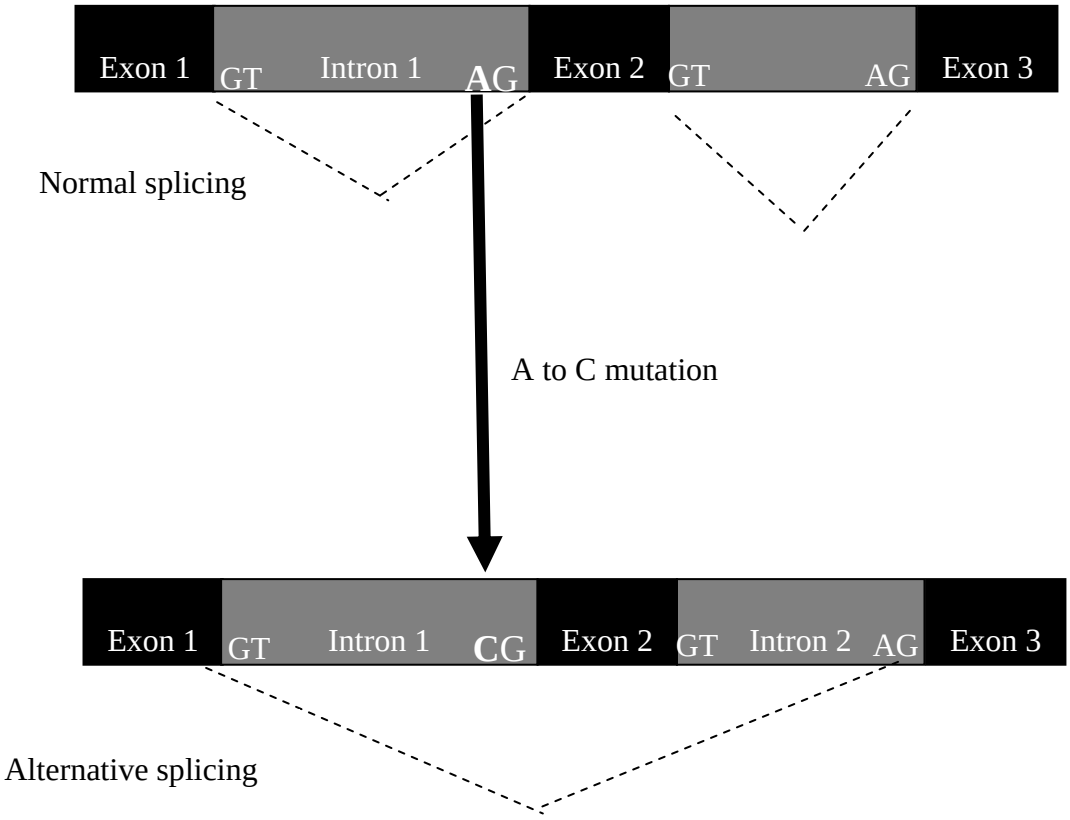

Figure 3. An illustration of exon skipping.

Splicing of an intron requires an essential signal: (GT.....AG). If the splice acceptor site AG is mutated e.g., A to C, the splicing machinery will look for the next acceptor site. As a result, the exon between two introns is also removed resulting in exon 2 being skipped.

A common feature of proteins that regulate splicing is the presence of two functional domains, an RNA-binding domain and a protein-protein interaction domain. The best characterized RNA-binding domains are the RNA recognition motif (RRM) and K-homology (KH) domains [33]. The three-dimensional structures of the two domains are distinct, as are the general features of the RNA-binding sites they recognize. Most ESEs are recognized by SR (serine/ arginine) proteins, which contain one or more RRM domains and an rich SR protein-protein interaction domain. SR proteins are essential, multifunctional splicing factors required at different steps of spliceosome assembly. They are also thought to mediate cross-intron interactions between splicing factors bound to the 5' and 3' splice sites. Finally, SR proteins are required for cross-exon interactions in both constitutively and alternatively spliced pre-mRNAs.

The role of regulatory SR proteins in splice-site selection can be affected by the promoter that generates the pre-mrna [33]. Thus transcription of the same pre-mRNA from different promoters can produce distinct mRNAs. This mechanism of alternative splicing could be a consequence of the coupling between transcription and pre-mRNA splicing. For example, particular SR proteins could be differentially recruited to RNA polymerase complexes assembled on different promoters, and then transferred to cognate splicing enhancers to promote the inclusion of specific exons.

1.3.4 Mechanism of alternative splicing

Majority of genes in eukaryotes undergo constitutive splicing in which each exon in a pre-mRNA is spliced with the most adjacent flanking exons. Hence only one mature mRNA is formed from a given pre-mRNA. However in alternative splicing, the exons

from a pre-mRNA can be ligated in different combinations resulting in the differential joining of 5' and 3' splice sites to produce multiple mature mRNAs differing in the precise combinations of their exon sequences. Transcripts are not always processed in the same way. For example, exons can be extended or shortened, skipped or included, and introns can be removed or retained in the mRNA. In some cases, exons are included in the mRNA in a mutually exclusive manner. These variations create multiple alternative splice isoforms of a gene sequence. One gene can encode multiple versions of mRNA, and each splice isoform of an mRNA holds instructions for making a different protein, thus attributing to protein diversity. Alternative splicing represents a means of producing two or more distinct but related proteins from the same gene. This process may occur in different tissues. The proteins translated from the transcript may have distinct functions.

It was estimated that at least 35% of human genes are alternatively spliced based on the analysis of Expressed Sequence Tags (EST) [19]. Current literature suggest that there is an average of three splice isoforms per gene [29] but the actual number of splice isoforms associated with each gene varies significantly. Alternative splicing occurs either at specific development stages or in different cell types. Alternative splicing has been implicated in various processes including sex determination [32] apoptosis [3,24], neuronal signaling [17] and acoustic tuning in the ear [54].

Although there are very few examples in which the mechanisms of alternative splicing are known in detail, a general outline has been established in Figure 4.

The mechanism of exon recognition in constitutively spliced pre-mRNAs provides the basis for positive and negative regulation of alternative splicing. The organization of

regulatory sequences within pre-mRNA (ESEs, ESSs, ISEs and ISSs) and the relative ratios of different regulatory proteins determine which splice sites are used in the splicing reaction. This, in turn, determines which exons are included in the mRNA.

Alternative splice-site selection in mammals is controlled by differential binding of regulatory proteins to splice sites, enhancers and silencer sequences within the pre-mRNA. The organization of these sequences and the interplay of different regulatory proteins determine the outcome of the splicing reaction.

1.3.5 Significance of alternative splicing

Alternative splicing of RNA is a common phenomenon in the regulation of eukaryotic gene expression. Study on this issue is of great importance to the research on molecular biology. Another important issue of concern is whether the alternatively spliced isoforms of a given nucleotide sequence can be correctly predicted. Having more than one protein per gene raises a critical issue for drug discovery in determining which isoform is the drug acting on and is it the correct isoform for the disease. Currently, the physiological effects of switching alternative splicing patterns and the results from the switch to be used as potential therapeutic agents are being investigated [52].

Abnormalities in the splicing process can lead to various disease states [16]. Splice variants with different functions have been linked to a variety of cancers, and genetic diseases such as thalassemia and cystic fibrosis [52]. Defects in the β-globin genes which cause β-thalassemias are caused by mutations in the sequences of the gene required for

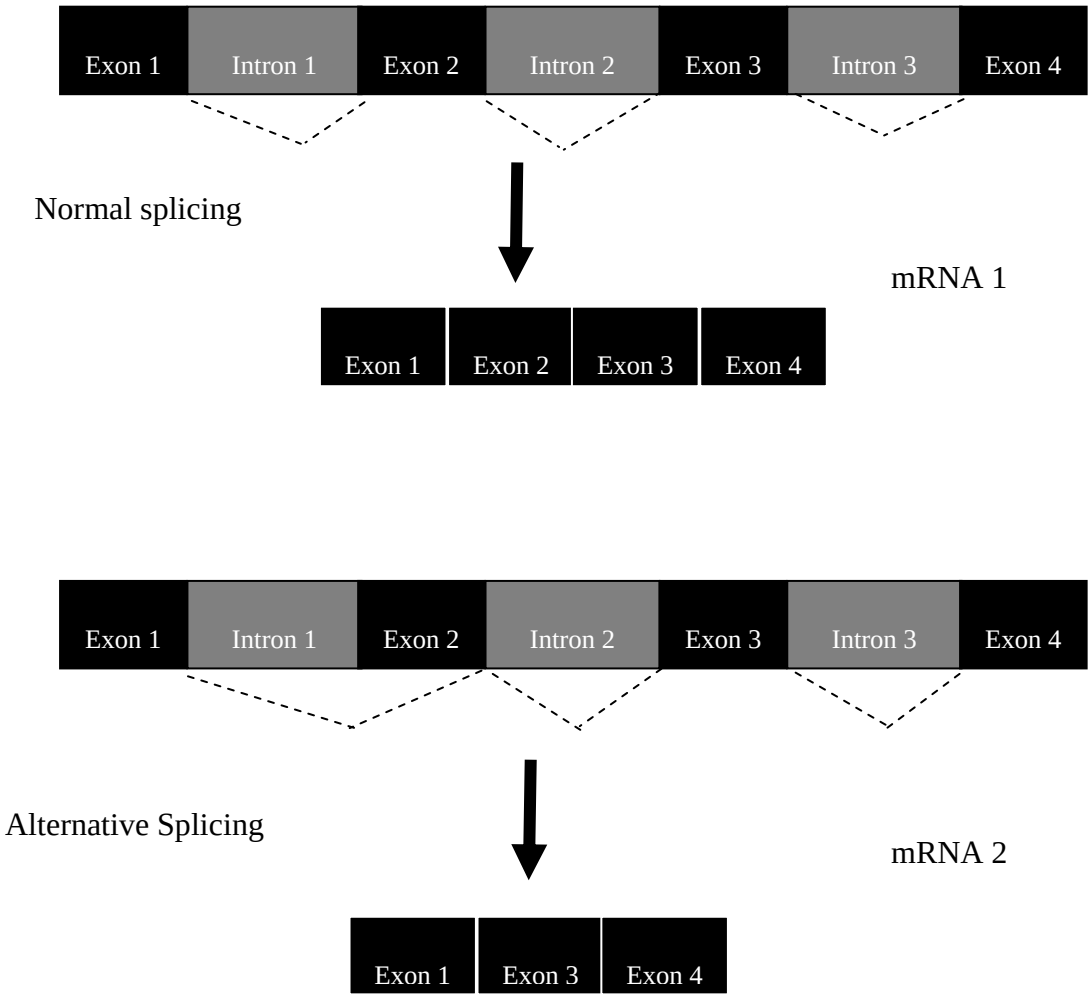

Figure 4. Alternative splicing produces variant proteins and splicing patterns.

During normal splicing, all the introns are spliced out, leaving the exons to remain in the mRNA. When alternative splicing occurs, both intron 1 and exon 2 are spliced out resulting in different mRNAs variants producing different proteins.

intron recognition and, therefore result in aberrant splicing of the β-globin primary transcript. Patients suffering from a number of different connective tissue disorders exhibit humoral auto-antibodies that recognize cellular RNA-protein complexes. Patients suffering from systemic lupus erythematosis have auto-antibodies that recognize the U1 RNA of the spliceosome. Aberrant splicing also plays a role in diseases such as Familial isolated growth hormone deficiency type II, Frasier syndrome, Frontotemporal dementia and Parkinsonism linked to Chromosome 17 (FTDP-17), Spinal muscular atrophy and Myotonic dystrophy [16]

Many cancer-associated genes are alternatively spliced. Although the functions of most of these splicing variants are not well defined, some have antagonistic activities related to regulated cell death mechanisms. In a number of cancers and cancer cell lines, the ratio of the splice variants is frequently shifted so that the anti-apoptotic splice variant predominates. This observation suggests that modification of splicing, which restores the proper ratio of alternatively spliced gene products, may reverse the malignant phenotype of the cells and offer a gene specific form of anticancer chemotherapy [10].

Several cancers and inherited diseases in humans are associated with mutations that cause unusual exon-skipping. Usually these mutations affect the splice sites. For example, a point mutation at the 5’ splice site of exon 7 of the Wilm’s tumor suppressor gene causes unusual skipping of exon 7 and generates a truncated protein that is associated with Wilm’s tumor. Mutations located outside traditional splice sites, either internally within the exon or in the flanking intron sequences have also been reported to be associated with

exon skipping and disease. For example, mutations in exon 18 of the breast cancer BRCA1 gene cause aberrant skipping of the entire constitutive exon.

Controlling how each cell responds to a diverse array of signals can be achieved through alternative splicing of its receptors and signal transduction molecules. A study revealed that alternative splicing occurs more frequently in molecules involved in metabolism, nucleic acid binding proteins and proteins involved in transcription and RNA processing [37]. Alternatively spliced genes also encode secreted proteins, signal transduction molecules and proteins involved in apoptosis. Majority of tissue specific expression occurs in the brain [50]. The brain has the largest total number of tissue specific splice forms, which represented 18% of all alternative splicing events observed [55].

1.3.6 Types of alternative splicing

A study [7] has generated and analyzed a high quality data set of EST and mRNA confirmed constitutive and alternatively spliced introns and exons for human genes. They have observed alternative events described as : (i) cryptic (or skipped) exons, where an entire alternative (or constitutive) exon is seen in some transcripts but not in others (i.e. a cassette exon); (ii) exon/intron isoforms, where use of alternative donor or acceptor splice sites leads to truncation/extension of exons/introns; and (iii) intron retention, where an intronic region is not spliced out.

Exon skipping is the most common of the alternative spliced forms (Figure 5). This pattern occurs in different tissues or at different developmental stages. The splicing variants of a gene cluster refer to all the distinct mRNA isoforms of that gene cluster. However, the number of mRNA isoforms might not be the same as the number of protein

variants because some mRNA isoforms give rise to the same protein variant. For example, alternative splicing may occur in the untranslated region of the gene where it could have an effect on mRNA stability or expression without a change in the protein product. In this study, we are only interested in alternative splicing that changes the coding sequences of the proteins.

1.3.7 Current methods for detecting alternative splicing

There is an enormous gap between the relatively small amount of detailed biochemical studies on alternative splicing (compared with the enormous amount of work on transcriptional regulation), and the recent rapid growth in alternative splicing data generated by high-throughput genomics studies. For example, tissue-specificities for only a small number of alternatively spliced genes are listed in current alternative splicing databases. The mechanism of alternative splicing regulation has been studied only for a small number of genes. A recent genomics study has identified 667 to 2873 tissue-specific alternative splice variants in human genes [55]. These high-throughput data could be useful for many biologists, leading to a rapid expansion in alternative splicing research with well-annotated datasets.

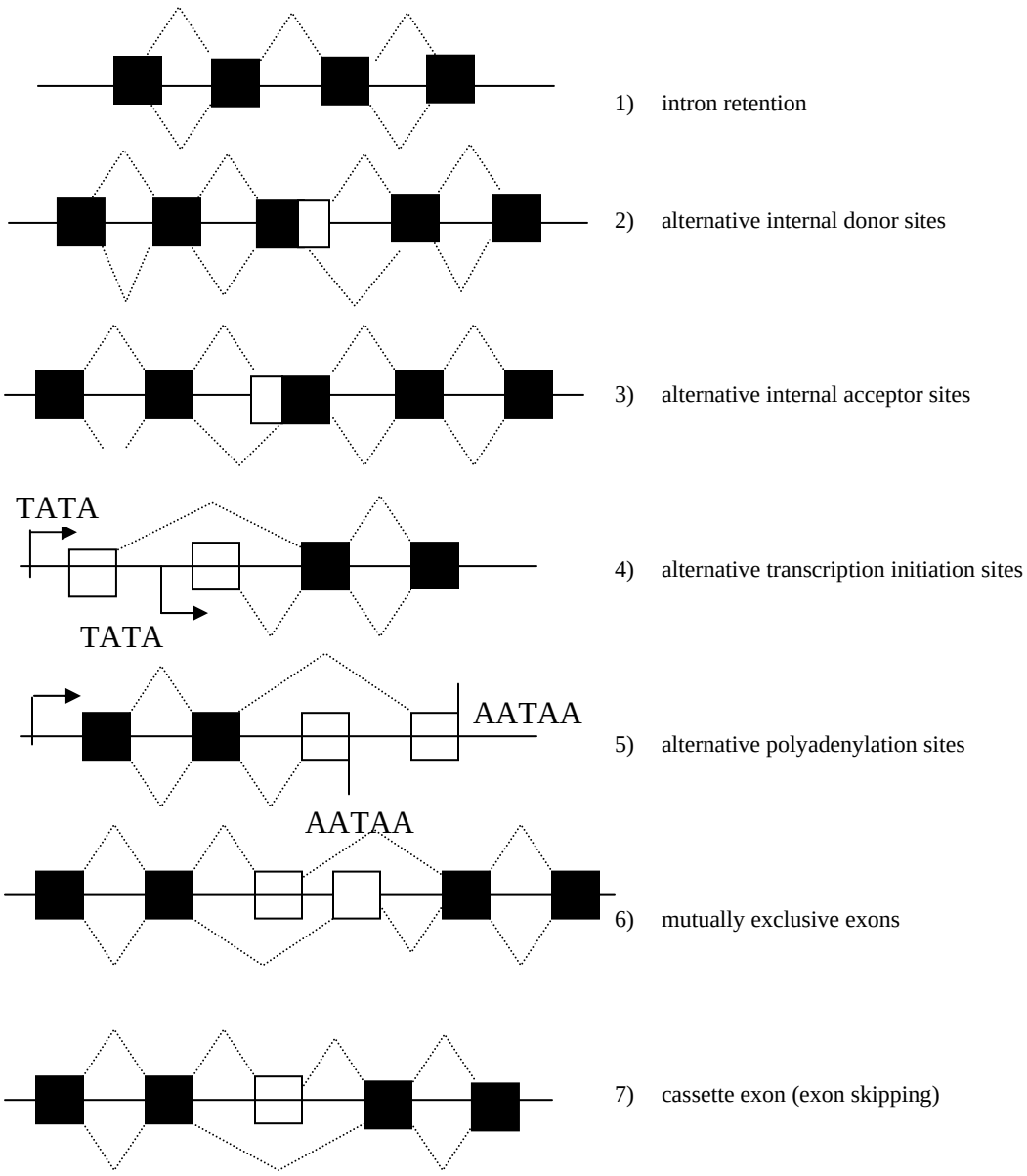

Figure 5. The various types of alternative splicing

(1) intron retention, (2) alternative internal donor sites, (3)alternative internal acceptor sites, (4) alternative transcription initiation sites, (5) alternative polyadenylation sites, (6) mutually exclusive exons and (7) cassette exon (exon skipping). Constitutive and facultative exons are shown as black and empty boxes, respectively. Introns are depicted as lines. AS is shown by the diagonal dashed lines above and below the gene. Transcription initiation and polyadenylation sites are denoted by TATA and AATAA.

In recent years, alternative splicing has been studied intensively in hundreds of human genes, and it appears to be widespread, occurring in 5 – 30% of human or perhaps as many as 35 – 40% [2,32,35,43,47,49]. It has been reported that alternative splicing can be detected in expressed sequence tag (EST) sequencing and genome wide detection has been performed [35,36]. Alternative splicing also has been identified in a collection of full-length mRNAs [2]. Putative alternative splicing information has been derived from the alignment of proteins, mRNA and EST data against human genomic DNA sequences [22]. Based on the estimates of the total number of human genes, it is likely that at least 10,000 – 20,000 human genes are alternatively spliced [15,31].

Large-scale analyses were conducted in the past few years by several independent groups using different methods [9,29,35,37]. The estimation of the proportion of human genes exhibiting alternative splicing is larger than previous expectations. The highest estimation is from International Human Genome Sequencing Consortium. They found that 59% of the genes on human chromosome 22 have at least two spliced variants. All the five groups [2,9,29,35] also pointed out that these figures are probably an underestimation, because (i) the EST database does not cover the entire repertoire of tissues or developmental states, and (ii) precautions taken to avoid false positives were extremely stringent hence resulting in lower number.

Bioinformatics plays an important role in studying alternative splicing. This includes (i) splice site prediction based on sequence comparison, (ii) search for regulatory elements by computational analysis and (iii) the construction of relational databases.

The increasing number of sequences from proteins, mRNA, cDNA and expressed sequence tags (ESTs) in the databases provide valuable evidence that can reveal splice variants. The use of computational alignment methods such as BLAST [1], BLAST-Like Alignment Tool (BLAT) [26] and SIM4 [18] allows the investigation of alternating splicing in the human genome. BLAT is used for nucleotide alignment between mRNA and genomic DNA taken from the same species. Sim4 is a computer program that aligns cDNA sequence with a genomic DNA sequence.

A number of techniques and databases have been developed for the study of alternative splicing. We summarize these techniques and databases in Table 1. Details are available in Appendix A.

Database/Algorithm	Data Source	No. of genes/spliced transcripts	Organism
ASDB	SWISS-PROT/ GenBank literature	1,922 protein and 2,486 DNA sequences	Human, Mouse, Rat, Fly, Worm, Chicken, Virus, Bovine and Rabbit
ASAP	mRNA–EST-genomic sequence alignment	7,991 genes/ 30,793 spliced transcripts	Human
SpliceDB	EST mapped to genomic sequence	43,337 splice junction pairs	Mammals
TAP	EST mapped to genomic sequence	1,124 transcripts	Human and mouse
STACKDB	Clustering of EST	270,515 clusters	Human
HASDB	mRNA and EST	6,201 spliced transcripts	Human
EASED	EST and mRNA	30,000 spliced transcripts	Human, Plant, Cow, Worm, Fly, Fish, Mouse, Rat and Frog
PALSDB	mRNA sequence in UniGene cluster and EST	19,936 (human) and 16,615 (mouse) UniGene clusters	Human and mouse
ProSplicer	Protein, mRNA, and EST sequences	21,786 genes	Human and mouse
ASD	EST/cDNA sequences and literature	8,314 genes	Human, fly, mouse, worm, rat, chicken, cow and zebrafish and plant
ASHESdb	Genomic/full-length cDNA sequences	1,229 genes/ 9073 spliced transcripts	Human

Table 1. The different alternative splicing databases developed by various research groups.

1.3.8 Problems and issues on existing AS detection methods

Current AS detection methods use EST resources to identify alternative splicing. The EST based computational approaches are more error prone, have limited gene coverage of transcript resources available and do not have confidence attached to the predicted events. EST data are potentially noisy as a result of sequencing errors, and may contain ambiguous and missing bases. Thus, it is not for sure if it is a reflection of a genuine genetic event resulting from a match to an alternative form when searching the sequence database with an EST query. Moreover, as EST sequences are partial sequences of gene transcripts, the alignment result may be incomplete and only show partial matching regions when they are compared to the genomic coding sequences [22]. Since ESTs cover a limited number of tissues and developmental or disease states there are likely to be many more spliced isoforms to discover. Furthermore, because ESTs are derived from sequencing the ends of cDNAs and covers only a portion of most cDNAs thus many of the splicing events that would affect the coding region of a gene will not be represented in EST libraries.

To circumvent these problems, our approach relies on the reliability of full-length cDNAs sequences to provide a more complete alignment to the human genome and improve the quality of AS detection.

Chapter 2

Material and Methods

In this study we used publicly available human full length cDNA sequences from numerous tissues to analyze the extent of alternative splicing in the human genome. We also developed a process to systematically identify and construct a database of putative alternatively spliced transcripts.

2.1 Material

The datasets used in our study consists of publicly available data from NCBI.

- Human full-length cDNA sequences from the Mammalian Gene Collection
- Human genome sequence Build 34.1 from NCBI

The Mammalian Gene Collection (MGC) Release 12 January 2004 contains 15,454 human full-length cDNA sequences in FASTA format [46]. The MGC's website provides the cDNA description and lists their tissue origin.

The NCBI Build 34.1 (ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapi_ens/) human genome data is comprised of GenBank flat files and FASTA formatted files (482 sequences from chromosome 1 – 22, X and Y). The GenBank annotations provided (as shown in Figure 6) include the chromosome location, origin, accession number, gene name, gene start and end position, exon start and end position.

There is more EST data than the full length cDNA available in public databases [28]. There are over three million human EST sequences clustered into ~96,000 gene clusters in UniGene as compared to 15,454 full length cDNA sequences from the Mammalian Gene Collection. Although we are likely to find more alternative splicing events using a larger dataset with ESTs, we chose to use full length cDNA data to get a higher quality database of alternative splice variants.

2.2 Methods

Exon skipping in transcript isoforms is a frequent event altering the protein coding sequence of genes [29,35]. The study of such events is of great importance in research on molecular biology because a complete understanding of all spliced variants of a gene will help in efficient gene discovery and target validation. Therefore, the identification of exon skipping patterns in the human genome data is particularly valuable in the identification of alternative spliced candidates.

We describe here a method which uses a computational process to identify alternative splicing. In this approach, we align cDNA data against human genomic sequences to find the occurrences of exon skipping. cDNA sequences have no introns and are reverse transcribed from a complete mRNA molecule representing the transcribed parts of the genome. We focus on finding the exons in the genomic sequences, which are omitted or skipped during the splicing process.

1: NT_011516. Homo sapiens chro...[gi:14776721]

```
LOCUS       NT_011516              234226 bp    DNA     linear   CON 19-FEB-2004
DEFINITION  Homo sapiens chromosome 22 genomic contig.
ACCESSION   NT_011516
VERSION     NT_011516.5  GI:14776721
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 234226)
  AUTHORS   International Human Genome Sequencing Consortium.
  TITLE     The DNA sequence of Homo sapiens
  JOURNAL   Unpublished (2003)
COMMENT     GENOME ANNOTATION REFSEQ:  Features on this sequence have been
            produced for build 34 version 3 of the NCBI's genome annotation
            [see documentation].
            On Jul 16, 2001 this sequence version replaced gi:13629036.
            The DNA sequence is part of the second release of the finished
            human reference genome. It was assembled from individual clone
            sequences by the Human Genome Sequencing Consortium in consultation
            with NCBI staff.
            COMPLETENESS: not full length.

FEATURES             Location/Qualifiers
     source          1..234226
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="22"
     source          1..37693
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /clone="c4G1"
                     /note="Accession AP000522 sequenced by Keio University
                     School of Medicine, Tokyo, JAPAN"
     source          37694..76727
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /clone="c60H5"
                     /note="Accession AP000523 sequenced by Keio University
     gene            2994..13236
                     /gene="LOC128939"
                     /note="Derived by automated computational analysis using
                     gene prediction method: GNOMON."
                     /db_xref="GeneID:128939"
                     /db_xref="InterimID:128939"
     mRNA            join(2994..3454,4682..4826,5999..6143,6486..6608,
                     6734..6800,12157..12316,12811..13236)
                     /gene="LOC128939"
                     /product="hypothetical LOC128939"
                     /note="Derived by automated computational analysis using
                     gene prediction method: GNOMON."
                     /transcript_id="XM_066243.2"
                     /db_xref="GI:37559466"
                     /db_xref="GeneID:128939"
                     /db_xref="InterimID:128939"
     CDS             join(2994..3454,4682..4826,5999..6143,6486..6608,
                     6734..6800,12157..12316,12811..12978)
                     /gene="LOC128939"
                     /codon_start=1
```

Figure 6. Genbank format file of chromosome 22 downloaded from NCBI.

This file contains information about the gene name, location and its exons' start and end position.

An overview of the pipeline process we used to identify AS by exon skipping is illustrated in Figure 7. The originality of this research lies in the AS identification method which utilizes full length cDNA as against commonly used EST based methods. The use of full length cDNA distinguishes this method from other available methods and thus helps to reduce false positive detections of AS. Subsequently, a user friendly web based relational database (POSTGRES) for alternatively spliced human genes is constructed. The data is available online for the scientific community.

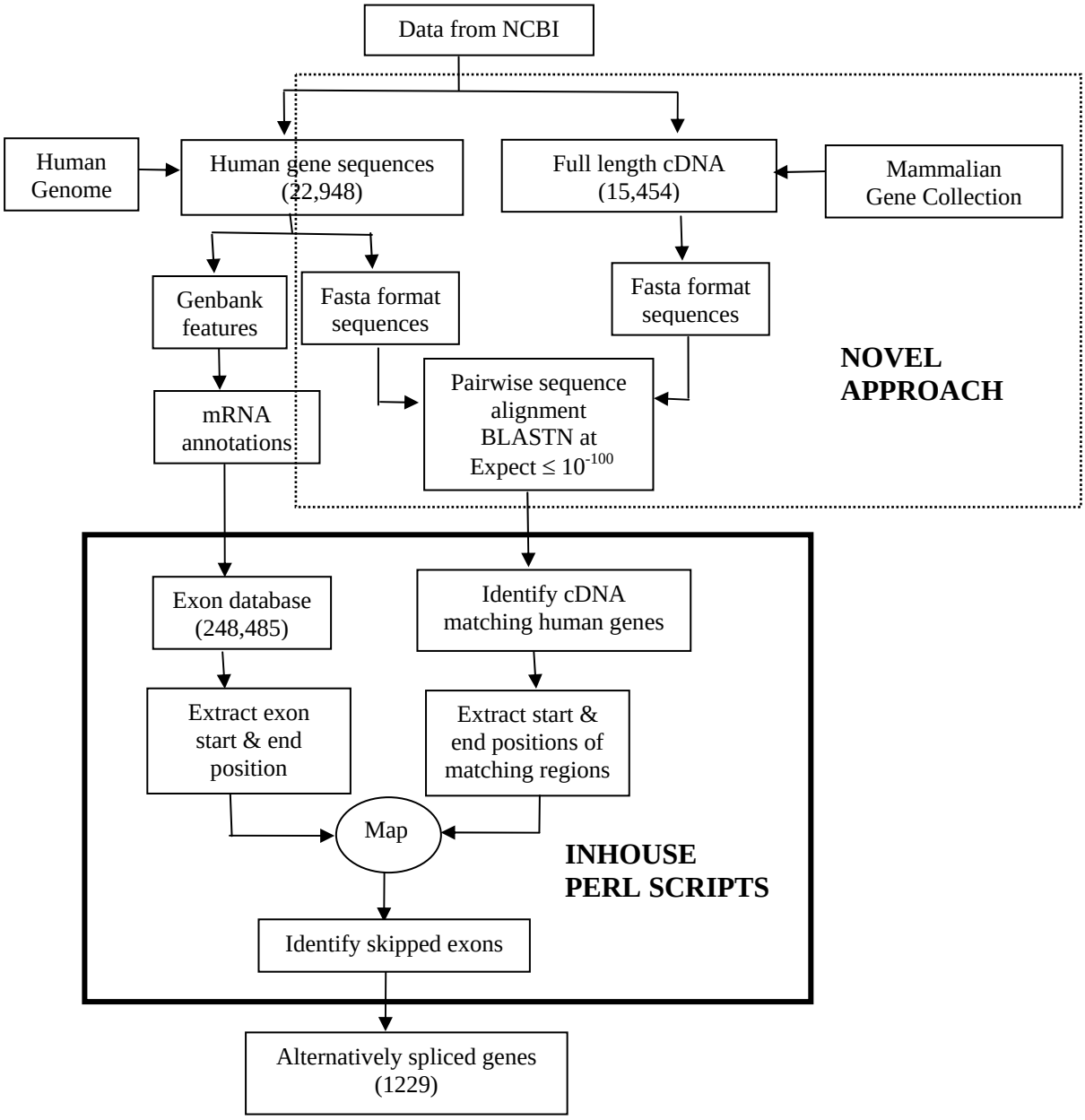

Figure 7. Flowchart for identifying AS in human

This flowchart summarizes the methodology used to identify alternatively spliced human genes by exon skipping (exons are skipped, for example exon 1 connects with exon 3 leaving exon 2). Our approach exclusively utilizes full length cDNA sequences to detect alternatively spliced genes (indicated by dotted border). This method reduces false positives compared to the methods that uses EST (expressed sequence tags). We developed in house PERL scripts for information extraction and mapping (indicated by thicker border).

2.2.1 Pairwise sequence alignment

The NCBI BLASTN (version 2.2.6) algorithm was used to align the 482 human chromosomal fragments against the 15,454 full length cDNA with a BLAST cut off of E $\leq 10^{-100}$. The program built an index of the query sequence and rapidly scanned for relatively short matches (hits), extended these into high-scoring pairs (HSPs), and then returned each area of homology between two sequences as separate alignments. An extension was triggered when one or two hits occurred in proximity to each other. This time consuming and intensive process took 2 weeks and was completed using distributed computing and was run on a 64 bit supercomputer at the Bioinformatics Institute, Singapore. The data and results generated took up 100 gigabytes of storage.

The BLAST results comprising of the set of alignments between cDNA and genomic sequences is shown in Figure 8, and the results were summarized using a PERL program we developed for easier subsequent processing into a file as shown in Table 2.

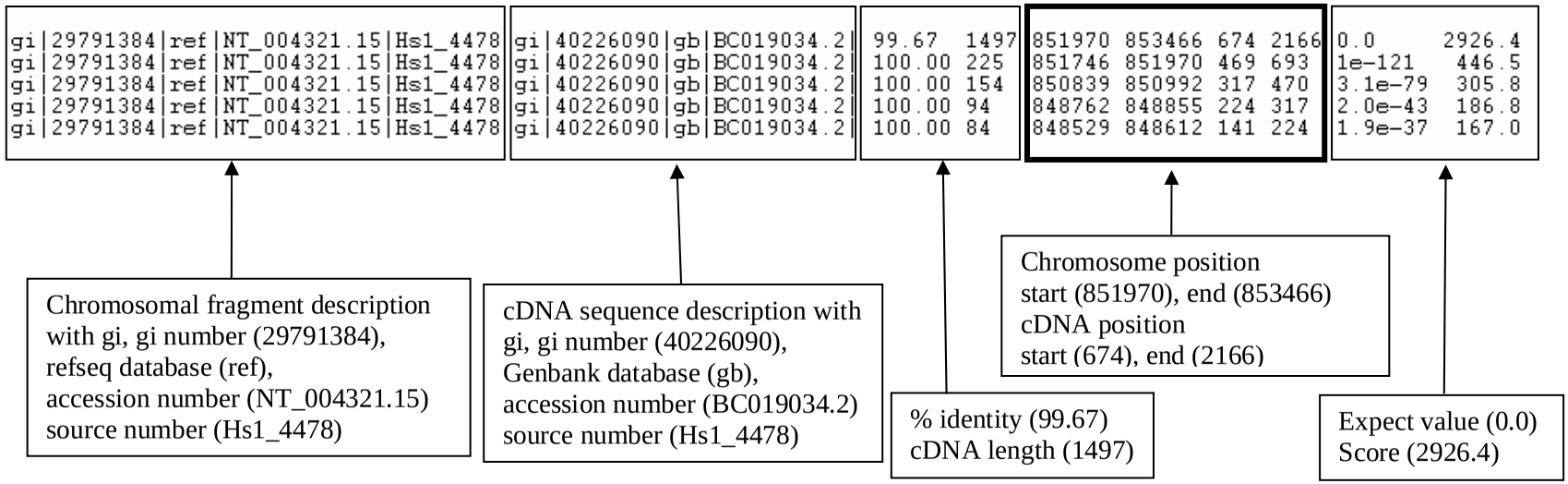

Figure 8. Sample BLASTN results from the pairwise alignment.

Information indicated in the box (highlighted with darker border) is the start and end positions of cDNA regions matching the chromosomal fragments extracted from the BlastN results. The first two columns within the highlighted box are the chromosome start and end positions, while the last two columns are the cDNA start and end positions.

chromosome fragment id	chromosome start position	chromosome end position	cDNA id	cDNA start position	cDNA end position
29791384	851970	853466	40226090	674	2166
29791384	851746	851970	40226090	469	693
29791384	850839	850992	40226090	317	470
29791384	848762	848855	40226090	224	317
29791384	848529	848612	40226090	141	224

Table 2. An illustration of BLASTN results summary.

The results were summarized into the chromosome and cDNA id with their start and end position for easier subsequent processing.

2.2.2 Creation of an exon database

The exon positions provided in the GenBank's annotations were predicted using Gnomon (Appendix B). We constructed an exon database by extracting the exon positions from the mRNA FEATURE annotations (rather than the CDS annotations, because full-length cDNA sequences contain not only the coding regions but also the 5' and 3' UTR) in the GenBank files. We automated the exons' positions and sequences extraction process using a program we developed (called exondb.pl) and stored them in a exon database. The program read the GenBank's feature information as shown in Figure 9, based on the following specification:

(1) Complement (location) which 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) as shown in Figure 9 implies that the indicated elements should be joined (placed end to end) to form one contiguous sequence.

Information on the exon positions were extracted from GenBank's annotation files and summarized into a format as shown in Table 3.

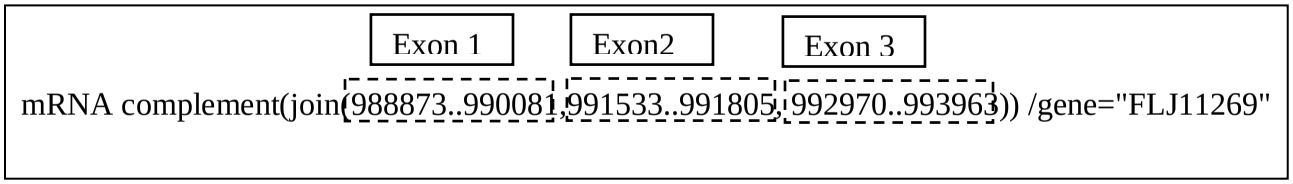

Figure 9. Information from the GenBank mRNA feature file.

Information provided shows the start and end positions for three exons belonging to the gene FLJ11269. Join implies that the indicated elements should be joined (placed end to end) to form one contiguous sequence.

chromosome fragment id	gene name	exon start position	exon end position
37547298	PEX10	313905	314935
37547298	PEX10	315585	315720
37547298	PEX10	315821	315996
37547298	PEX10	317553	317959

Table 3. Exon database summary.

Summary of the exon start and end positions extracted from GenBank mRNA feature for gene PEX10.

2.2.3 Identification of exons matching cDNA

We developed a PERL script (called matchexon.pl) to automatically identify exons matching the cDNA. We used the alignment summary to generate regions in the genomic sequences that match the cDNA sequences. Thus, the exon position matching the cDNA was determined. Consequently, the exon number matching the cDNA was identified using the exon database. A high level of stringency was applied at this stage, to allow a maximum error of 10 nucleotides in the start and end positions of the matches against the exons.

2.2.4 Identification of exon skipping patterns

We also developed a PERL script (called find_exonskip.pl) to automatically identify exon skipping patterns from the list of different exons matching the cDNAs . Exons which match a cDNA were represented as class “1” and “0” for those which did not (as shown in Table 4). Alternatively spliced genes were predicted based on the exons with more than one cDNA match.

2.2.5 Data analysis

Further analysis was carried out to determine the significance of the predicted AS based on their exon skipping patterns and tissue specificity.

Chromosomal fragment ID	Gene name	Exon start position	0 if exon is skipped, 1 if exon is not (cDNA ID if exon is not skipped)
37547298	PEX10	37588974	0111111
37547298	PEX10	17390442	0110110

Table 4. Summary of exon to cDNA match.

Information provided contains the start positions of exons 2 and 3 from gene PEX10 matching cDNA id 37547298. Exon 1 which does not match any cDNA is denoted as "0". Information provided shows that exons 1, 4 and 7 in gene PEX10 are skipped.

2.3 Database schema

Information on the predicted alternatively spliced genes and their exon skipping patterns are stored in a set of five SQL tables (schema shown in Figure 10). The five SQL tables consist of

- Chromosome table (chr_table) which has information on the chromosome fragment id (chr_frag_id), chromosome number (chr_name), GenBank accession number (genbank_id) and organism it is derived from.

- Gene table (gene_table) which has information on the chromosome fragment id (chr_frag_id), exons, GenBank accession number (genbank_id), gene name (gene_name), description (gene_desc) and tags to indicate alternative splicing event.

- Exon table (exon_table) which has information on the start (start_pos) and end (end_pos) position, exon number (exon_num) and gene name (gene_name).

- cDNA table (cdna_table) which has information on the cDNA accession number (cdna_id), name (cdna_name), exons that mapped to it (exon_map), name (gene_name) and tissue library where the cDNA is obtained.

- Exon sequence table (exonseq_table) which has information on chromosome fragment id (chr_frag_id), exon number (exon_num), gene name (gene_name) and exon sequences (exon_seq).

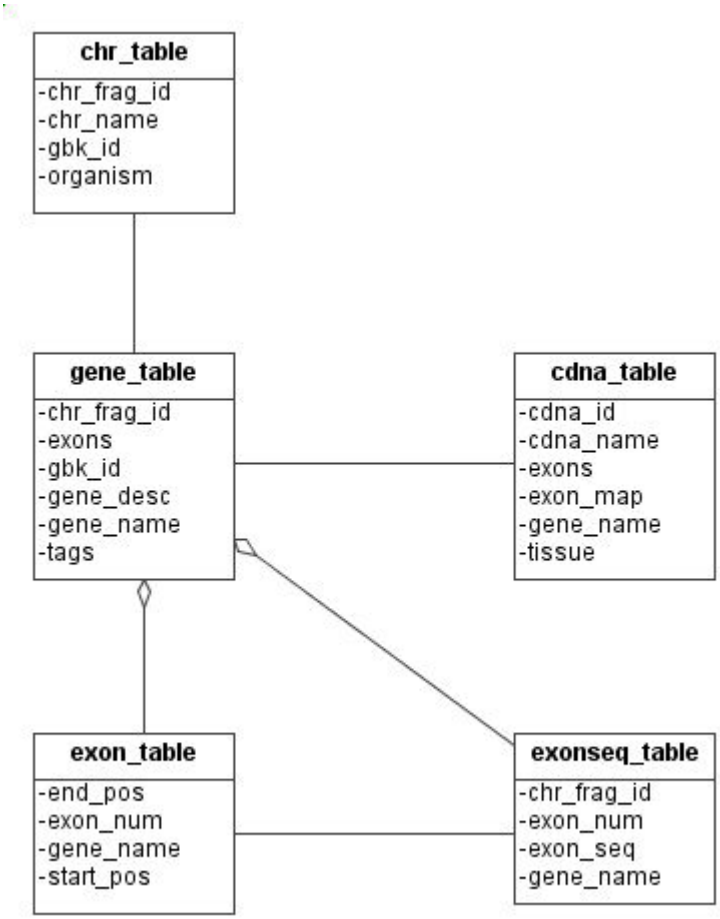

Figure 10. ASHES database schema.

The database includes five tables containing chromosome, cDNA, gene and exon information such as Chromosome table (chr_table), chromosome fragment id (chr_frag_id), chromosome number (chr_name), GenBank accession number (genbank_id), gene description (gene_desc), exon start (start_pos) and end (end_pos) position, exon number (exon_num), cDNA accession number (cdna_id), cDNA name (cdna_name), exons that mapped to it (exon_map) and Exon sequence table (exonseq_table).

2.4 Web interface and search engine

The database is made available to users via a web interface known as the Alternative Spliced Human Genes by Exon Skipping (ASHES) database.

The database is accessible via http://sege.ntu.edu.sg/wester/ashes/.

Chapter 3

Results and Discussion

3.1 Results

3.1.1 Genome-wide detection of alternative splicing

Our computational approach using full length cDNA to predict alternative splicing in the human genome found 1,229 genes, 2717 splice variants and stored them in a relational database (ASHESdb). In our study, 110,169,658 matches, from the pairwise alignment between human cDNA and genomic sequences met the BLAST cut off of $E \leq 10^{-100}$. They were summarized and mapped to the 248,485 human exons database extracted from Genbank annotations to predict exon skipping events. 1,229 human genes had exon skipping occurrences and were predicted to be alternatively spliced using our computational process. The exon skipping patterns from these genes were grouped into 1,055 types based on the different combinations of exons matching the same cDNA. The results were stored in a database which consisted of five SQL tables. The database is made available to users via a web interface developed using Perl CGI and HTML. It is designed to provide the user the ability to query the database via a range of database fields as shown in Figure 11.

The search engine has a simple query form, which allows users to search for exon skipping occurrences via a CGI written in PERL either by chromosome number, gene name or the exon skipping pattern. Users may also query the database using the BLAST function from http://sege.ntu.edu.sg/blast/blastaltspl.html.

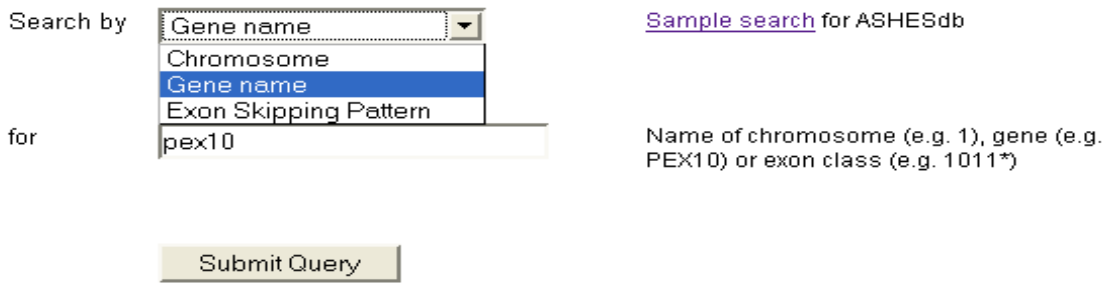

Figure 11. Input parameters for users to search ASHESdb.

Users may search the database by chromosome number, gene name and exon skipping pattern. The exon is denoted as 1 when it matches the cDNA and 0 when it is skipped. Pattern 1011 means the second exon is skipped out of 4 exons present.

Figure 12 shows how AS information is obtained from ASHESdb. The search result for gene name PEX10 returns a description of the gene, information about its exon skipping patterns, the two cDNA it matches to, the cDNA origin, description and GenBank cross references. Exon sequences are available for download and to be submitted to BLAST.

3.1.2 Tissue specific study of alternative splicing

Further analysis on the spliced variants predicted showed that majority of the alternatively spliced human genes match cDNAs from the brain as shown in Figure 13.

3.1.3 Exon skipping patterns analysis

Analysis based on the exon skipping patterns indicated that majority of the alternative spliced genes in our database have two variants (81%), followed by three variants (12%) as shown in Figure 14.

Table 5 shows that exon skipping occurs more frequently in first and last (terminal) exons based on calculations done to find the percentage occurrence of terminal exons versus non terminal exon skipping occurrences.

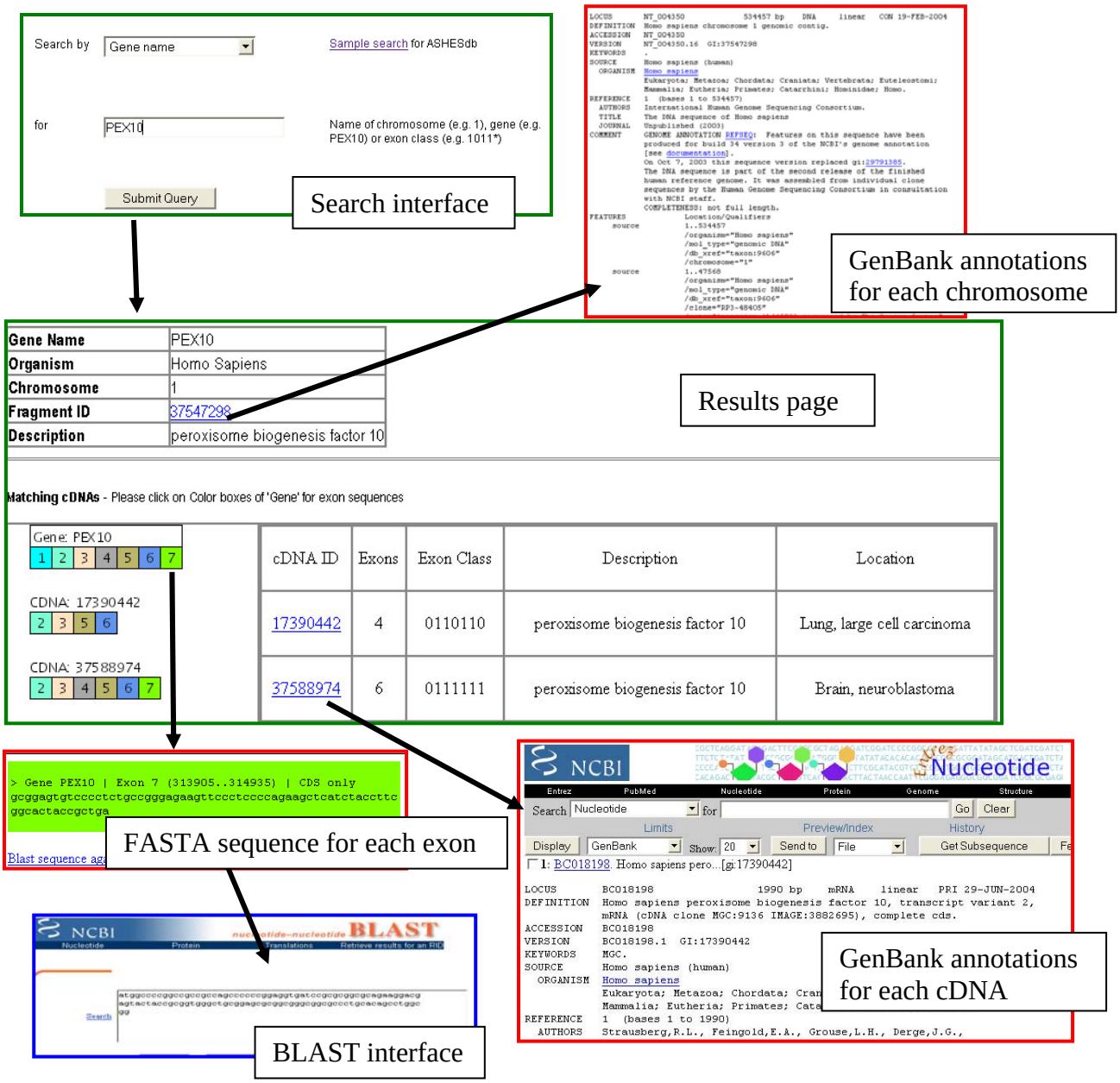

Figure 12. Search results for "PEX10" gene on chromosome 1.

The chromosomal fragment's Genbank ID, 37547298 is cross-referenced to the Genbank. The exon skipping pattern displayed on the results page also indicates which exons on the gene matches the cDNA. Information on the gene location on the chromosome, the number of exons matching the cDNA and the exon skipping class are also provided. Users may click on the exons to view their positions and to retrieve their sequences in FASTA format.

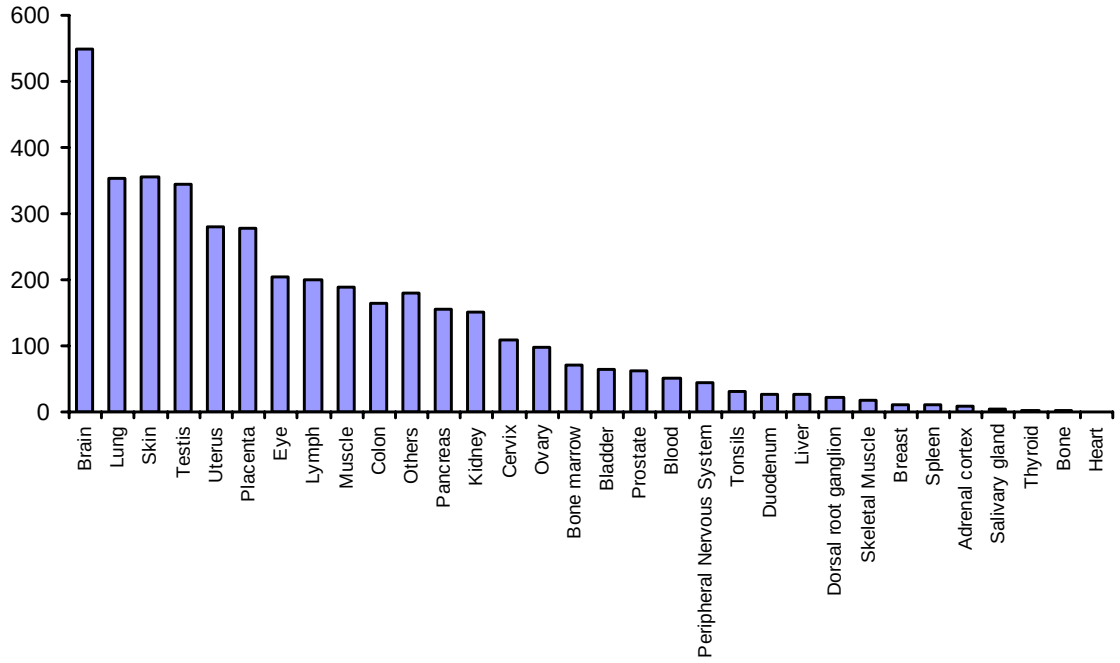

Figure 13 Distribution of genes exhibiting AS in different tissues.

The graph shows the highest prevalence of alternative splicing in the brain suggesting that AS play an important role in the nervous system whose functions require precise control of cellular differentiation, activation and to process large amounts of information.

Description	Number	% probability
Total Exon skipping patterns	1055	
Non-terminal exon skipping occurrences	708	67.1
Terminal exon skipping occurrences	827	78.4

Table 5. Exon skipping occurrence analysis.

Exon skipping occurs more frequently in terminal exons versus non terminal exon skipping occurrences. Terminal exons refer to the first and last exons in the mRNA.

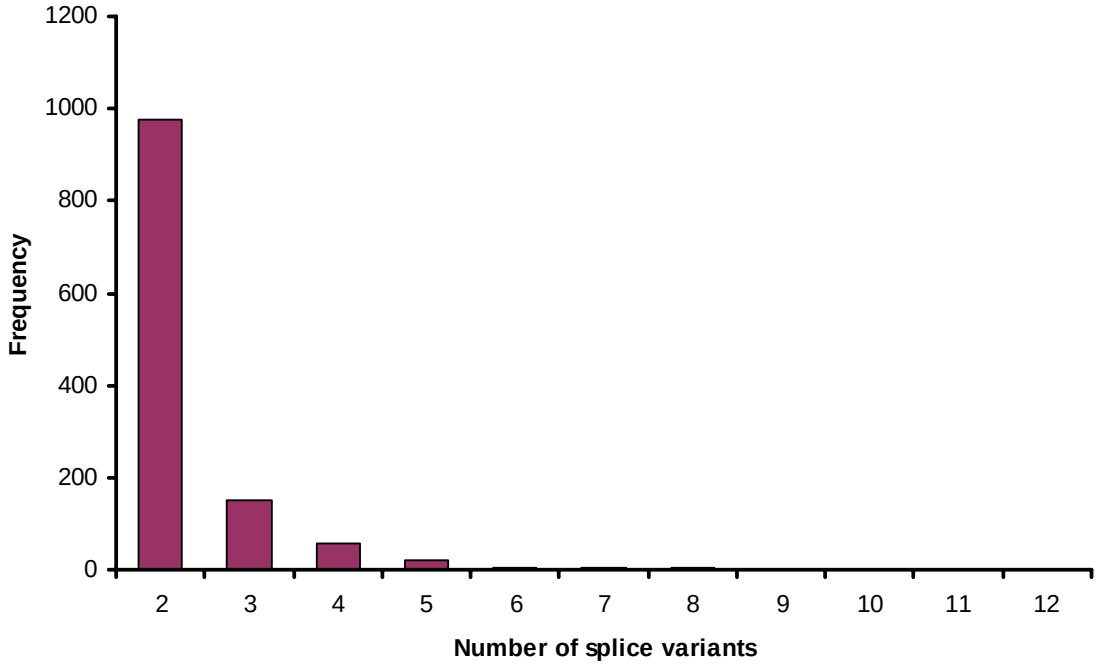

Figure 14 Number of distinct splice variants for each gene in the database.

Majority of the genes have two splice variants, hence they probably do not have many isoforms.

3.2 Discussion

Data collection is a crucial initial step during the construction of any database. Our results provide biologists with a web based relational human alternative splicing database and their exon skipping patterns. Using a computational process, our approach enables large-scale genome wide detection of AS. The use of full length cDNA helps to overcome the limitations of using EST data in AS prediction i.e. error prone and lack of confidence due to the limited transcript coverage.

Our AS prediction method is dependent on various factors. The 1,229 AS cases identified in our approach are with respect to the Gnomon's gene predictions for complete exon/intron structures of genes in genomic DNA provided by NCBI. Hence some of the identified exon skipping cases may also be false positives, if the skipped exons were false positives. Other groups developed their own splice site detection method and identified 796 entries containing alternative splicing events [7]. However, given the poor understanding of the biological rules of splice site selection, it will take some time before it is possible to predict all the alternatively spliced mRNAs that might derive from a given genomic locus. Programs robust enough to recognize cryptic consensus splice sites that are not normally used by the splicing machinery but recognize non-consensus splice sites that are actually used by the splicing machinery are still lacking. However, the discovery of other cis-acting sequences in the pre-mRNA, the exonic splicing enhancers (ESE) and inhibitor sequences will help to improve gene prediction algorithms.

Our computational approach uses the standalone NCBI BLAST for the pairwise alignment instead of other more sensitive algorithms such as FASTA and Smith

Waterman because of its speed, and required less time to complete this computing intensive process. The BLAST cut off of $E \leq 10^{-100}$ was set to ensure that only high scoring assemblies per gene were retained. A larger allowed error would have resulted in more matches and consequently more alternative splicing entries in our database, but the accuracy of the entries would have been compromised.

A stringent criteria was used in the identification of genes that have matches to two or more cDNAs with different exon skipping classes. Such genes are definite cases of alternative splicing as false positives are filtered out by the observation of the skipped exons in other cDNA matches.

Based on only chromosome 22, Hide et al. estimated that 5% the proportion of all transcripts corresponding to multi-exon genes exhibit exon skipping [20]. Our study also showed a similar estimation (5.5%) based on the 1,229 AS genes found out of 22,143 genes extracted from NCBI's human genome dataset. The current full length cDNAs in the Mammalian Gene Collection (MGC) consists of ~52% of all genes but it will grow to 67% in the near future (Collins, 2002). The collection grew from 9,000 to 15,454 in 1.5 years and the number is constantly increasing. As the cDNA set that was used for the search is not representative of all cDNAs found in humans, the 1,229 observed cases of alternative splicing are not exhaustive. However, the method can be extended to more ful length cDNA as they become available to build a complete human alternative splicing database.

Alternative splicing is found in various tissues. However, the tissue specificity distribution in Figure 13 indicates that our result is consistent with a similar study by Xu

et al, [55] based on EST data, which revealed that the brain has the largest number of tissue specific splice forms. Majority of tissue specific exon expression have been reported in the brain [45]. Previous studies also indicated that the nervous and immune systems are major loci of alternative splicing [42,48]. Differential inclusion or skipping of the variable alternatively spliced exon (VASE) in the gene for neural cell adhesion molecule (NCAM) in rat brain represses or promotes axon outgrowth during development.

In the nervous system, proteins involved in the formation of neuronal connections during development and proteins that mediate cell signaling show particularly high levels of molecular diversity created by alternative splicing. The Drosophila homologue of Downs syndrome cell adhesion molecule (Dscam) is a good example to illustrate the importance of AS in providing the intricate molecular diversity required for the specificity of cell signaling and communication [44]. The Dscam gene is involved in axon guidance in the developing brain. It contains 115 exons, 95 of which are alternatively spliced, and can potentially generate 38 016 isoforms. However, this is not common as majority of genes in our prediction (as shown in Figure 14) have two and three splice variants.

The prevalence of alternative splicing occurrences in the human brain tissue thus suggests that AS may play an important role in the

- nervous system whose functions require precise control of cellular differentiation, activation and to process large amounts of information

- formation of neuronal connections during development and mediate cell signaling and communication

Alternative splicing in the human brain tissue regulates the expression of a divergent set of spliced isoforms needed for its functional complexity. Tissue-regulated AS may play an important role in the differentiation of the different tissues. Thus, a more comprehensive study of AS require the integration of additional types of data upstream and downstream of splicing-factor expression. Upstream, splicing factors may be differentially regulated in different tissues or in response to different stimuli at the level of transcription, splicing, or translation, and are frequently regulated by post-translational modifications such as phosphorylation, so systematic measurements of splicing factor levels and activities will be required. Downstream, AS may affect the stability of alternative transcripts, and frequently alters functional properties of the encoded proteins, so systematic measurements of AS transcript and protein isoforms and functional assays will also be needed to fully understand the regulatory consequences of AS events

Alternative splicing may occur in the 3’ or 5’ untranslated regions (UTRs) or in the protein coding sequence. The 5’UTR sequence contains regulatory regions that control protein expression. Insertion or deletion of these regions will have a consequence on protein expression. The 3’ UTR region contains mRNA stability domains. Insertion or deletion of these domains has consequences on mRNA stability and therefore protein expression. Alternative splicing within the protein coding sequence results in altered protein structure and function. The pattern in table 5 presented that the terminal exons are more frequently skipped to minimize effects of protein structural and functional changes.

Terminal exons, being larger probably inhibit spliceosome formation and affect splice site recognition. This suggests the significance of exon size in splicing mechanism. To better understand this, we need more detailed analysis on

- the determination of exons conservation in the alternative spliced genes
- the role played by exon/intron architecture and length in splicing machinery and its effects on alternative splicing.

ASHESdb is a user friendly searchable AS database created using computational approach. It has a web interface to allow users to query data using keywords and accession numbers and sequence similarity search using BLAST. It provides information on genome wide detection of human alternative splicing including

- cross-references to literature and functional annotations,
- tissue specificity, protein isoforms and exon sequences for probe design
- graphical representations of splicing patterns.

Although various groups have constructed databases containing products of alternatively spliced gene products, it is difficult for users to judge the quality of prediction of the splice variants without experimental validation and literature support. ASHESdb helps researchers to detect novel isoforms and extract exon sequences for probe design. With the availability of full length cDNA and genome in other model organisms such as mouse

and zebrafish, it is also possible to carry out comparative genomics to study the homology of alternatively spliced products between human and other organisms.

Chapter 4

Conclusion

In conclusion, we have developed a computational pipeline process for genome wide detection of human alternative splicing. Using this approach, we identified 1,229 human genes and 2717 splice variants exhibiting AS based exclusively on full length cDNA. This method significantly reduces false positives introduced by procedures determined using EST data. Subsequently, a web based relational database (ASHESDB) is constructed to store and search these genes for public convenient access.

Users may use the database to search for alternative splicing candidates in the human genome, view the various types of exon skipping patterns and use the exon sequences to design probes for experimental validation.

Our results are consistent with that of work done by earlier reports. However, it is important to note that not all exon skipping cases in our database may be real. There may be cases of apparent exon skipping due to the incorrect prediction of exons in the draft human chromosomal sequence. However, the prediction programs have been optimized considerably. Further analysis is required with a more complete cDNA dataset to build a comprehensive AS database. A comprehensive database with multiple features is of value in the design of drug targets during drug discovery. It should be noted that the 2717 AS variants identified in this study are limited. This limitation could be overcome with the availability of more full length cDNA.

Alternative splicing is a complex yet important mechanism for generating diverse protein isoforms in human cells. A collection of full length cDNA sequences enables us to identify such variants. However, the combinatorially large possibility of splicing is the bottleneck for the complete understanding of this process.

References

1. Altschul, S.F., Miller, G.W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403 – 410.

2. Brett, D., Hanke, J., Lehmann G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Letters*, 474, 83-86.

3. Boise, L.H., Gonzalez-Garcia, M., Postema, C.E., Ding, L., Lindsten, T., Turka, L.A., Map, X., Nunez, G., and Thompson, C.B. (1993). Bcl-x, a Bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell*, 74, 597-608.

4. Burge, C.B., Tuschl, T. and Sharp, P.A. (1998). Splicing of precursors to mRNAs by the spliceosomes. *The RNA World*, 2, 525-560.

5. Burset, M., Seledtsov, I., and Solovyev, V. (2001). SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Research*, 29,255-259.

6. Caceres, J. and Komblihtt, A. (2002). Alternative splicing: multiple control mechanism and involvement in human diseases. *Trends in Genetics*, 18, 186-193.

7. Clark, F. and Thanaraj, T.A. (2002). Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Human Molecular Genetics*, 11, 451-464.

8. Collins, F. (2002). Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci U S A*, 26, 16899-16903.

9. Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P. and Mattick, J.S. (2000). ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genetics*, 24, 340-341.

10. Danielle, M. and Ryszard, K. (2000). Modification of alternative splicing pathways as a potential approach to chemotherapy. *Pharmacology and Therapeutics*, 85, 237-243.

11. Deutsch, M. and Long, M. (1999). Intron-Exon structures of eukaryotic model organisms. *Nucleic Acids Research*, 27, 3219 – 3228.

12. Dietrich, R., Incorvaia, R. and Padgett, R. (1997). Terminal intron dinucleotides do not distinguish between U2- and U12-dependent introns. *Molecular Cell*, 1, 151-160.

13. Dralyuk., I., Brudno, M., Gelfand, M., Zorn, M. and Dubchak, I. (2000). ASDB : Database of alternatively spliced genes. *Nucleic Acids Research*, 28, 297-297.

14. Duret, L., Mouchiroud, D. and Gautier, C. (1995). Statistical analysis of vertebrate sequences reveals that long genes are scarce in CG-rich isochores. *Journal of Molecular Evolution*, 40, 308-317.

15. Ewing, B. and Green, P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genetics*, 25, 232-234.

16. Faustino, N.A. and Cooper, T.A. (2003). Pre-mRNA splicing and human disease.*Genes and Development*, 17, 419-37.

17. Fettiplace, R. and Fuchs, P.A. (1999). Mechanism of hair cell tuning. *Annual Reviews of Physiol*ogy, 61, 809 – 834.

18. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. and Miller, W. (1998). A computer program for aligning cDNA sequence with a genomic DNA sequence. *Genome Research*, 8, 967-974.

19. Gravely, B. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics*, 17,100-107.

20. Hide, W., Babenko, A., Vladimir, N., Heusden, P., Seoighe, C. and Kelso, J. (2001). The contribution of exon-skipping events on Chromosome 22 to Protein Coding Diversity. *Genome Research*,11,1848-1853.

21. Huang, Y.H., Chen, Y.T., Lai, J.J., Yang, S.T. and Yang, U.C. (2002). PALS db : Putative alternative splicing database. *Nucleic Acids Research*, 30, 186 – 190.

22. Huang, H., Horng, J., Lee, C. and Liu, B. (2003). ProSplicer: a database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data, *Genome Biology*, 4**,** R29.

23. Ji, H., Zhou, Q., Wen, F., Xia, H., Lu, X. and Li, Y. (2001). AsMamDB : An Alternative Splice Database of Mammals. *Nucleic Acids Research*, 29, 260-263.

24. Jiang, Z..H.. and Wu, J.Y. (1999). Alternative splicing and programmed cell death. *Proc Soc Exp Biol Med*. Feb; 220, 64-72.

25. Kan, Z., Rouchka, E.C., Gish, W.R. and States, D.J. (2001). Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Research*, 11, 889-900.

26. Kent, J. W. (2002). BLAT – The BLAST –Like Alignment Tool. *Genome Research*, 12, 656 – 664.

27. Kent, W. J, and Zahler, A.M. (2000). The intronerator: exploring introns and alternative splicing in *Caenorhabditis elegans*. *Nucleic Acids Research*, 28, 91-93.

28. Kristiansen, T.Z and Pandey, A. (2002). Resources for full-length cDNAs. *Trends in Biochemical Sciences*, 27, 266-267.

29. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C. and Zody, M.C. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860-921.

30. Lee, C., Levan, A., Modrek, B. and Yi, X. (2003). The Alternative Splicing Annotation Project. *Nucleic Acids Research*, 31, 101–105.

31. Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L. and Quackenbush, J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genetics*, 25, 239-240.

32. Lopez, A.J. (1998). Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annual Review of Genetics*, 32, 279-305.

33. Maniatis, T. and Tasic, B. (2002). Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, 418, 236-43.

34. Miriami, E., Margalitt, H. and Sperling, R. (2003). Conserved sequence elements associated with exon skipping. *Nucleic Acids Research,* 31, 1974-1983.

35. Mironov, A.A., Fickett, J.W. and Gelfand, M.S. (1999). Frequent alternative splicing of human genes. *Genome Research*, 9, 1288-1293.

36. Modrek, B. and Lee C. (2002). A genomic view of alternative splicing. *Nature Genetics*, 30, 13-19.

37. Modrek B., Resch A., Grasso C., and Lee C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research*, 29, 2850 – 2859.

38. Moore, M.J., Query, C.C. and Sharp, P.A. (1993). Splicing of precursors to mRNA by the spliceosome. *The RNA World*, 2, 303-357.

39. Pollastro, P., Rampone, S. (2002). HS3D, a Dataset of Homo Sapiens Splice Regions, and its Extraction Procedure from a Major Public Database, *International Journal of Modern Physics* C, 13, 1105-1117.

40. Pospisil, H., Herrmann, A., Bortfeldt, R.H. and Reich, J.G. (2004). EASED: Extended Alternatively Spliced EST Database. *Nucleic Acids Research* Database issue.

41. Pruitt, K.D. and Maglott, D.R. (2001). RefSeq and LocusLink : NCBI gene centered resources. *Nucleic Acids Research*, 29, 137 – 140.

42. Seya,T., Hirano,A., Matsumoto,M., Nomura,M. and Ueda,S. (1999). Human membrane cofactor protein (MCP, CD46): multiple isoforms and functions. *International Journal of Biochemistry and Cell Biol*ogy, 31, 1255-1260.

43. Smith,C.W.J. and Valcarcel,J. (2000). Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends in Biochemica. Science*, 25, 381-388.

44. Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E.and Zipursky, S.L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101, 671-684.

45. Stamm,S., Zhu,J., Nakai,K., Stoilov,P., Stoss,O. and Zhang,M.Q. (2000). An alternative-exon database and its statistical analysis. *DNA Cell Biology*, 19, 739–756.

46. Strausberg, R.L., Feingold, E.A., Klausner, R.D. and Collins, F.S. (1999). The Mammalian Gene Collection. *Science*, 286, 455–457.

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

48. Smith., C.W.J. and Valcarcel, J. (2000). Alternative pre-mrRNA splicing : the logic of combinatorial control. *Trends in Biochemical Sciences,* 25, 381-388.

49. Sutcliff, J.D. and Milner, R.J. (1998). Alternative mRNA Splicing : The Shaker gene. *Trends in Genetics*, 4, 297-299.

50. Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O. and Zhang, M. (2000). An Alternative-Exon Database and Its Statistical Analysis. DNA and Cell Biology, 19, 739–756.

51. Sterner,D. A., Carlo T. and Berget S.B. (1996). Architectural limits on split genes, *Proc Natl Acad Sci U S A Biochemistry,* 93, 15081-15085.

52. Sazani, P. and Kole, R. (2003). Therapeutic potential of antisense oligonucleotides as modulators of alternative splicing. *Journal of Clinical Investigations*, 112, 481–486.

53. Thanaraj, T.A., Stamm, S., Clark, F., Riethoven, J.J., Le Texier, V. and Muilu, J. (2004). ASD: the Alternative Splicing Database. *Nucleic Acids Research*, 32, 64 – 69.

54. Ullrich, B., Ushkaryov, Y.A. and Sudhof T.C. (1995). Cartography of neurexins : more than 1000 isoforms generated by alternative splicing and expressed in distinct subsets of neurons. *Neuron,* 14, 197 – 507.

55. Xu, Q., Modrek, B.and Lee, C. (2002).Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Research*, 30, 3754 – 3766.

Appendix A – Alternative Splicing Databases

1. AsMamDB

AsMamDB contains information about alternative splicing in several mammals including alternative splicing patterns, gene structures, locations in chromosomes, products of genes and tissues in which they express [23]. This was generated before the completion of the human genome. Data for AsMamDB are collected from Genbank and Unigene. All Genbank entries containing the term "*Homo sapiens*" were used to create the human subset while entries containing the terms "*Mus musculus*" and "*Rattus norvegicus*" were selected to create the mouse and rat subsets, respectively. The alternative splicing patterns of each gene are depicted by an alignment of transcripts. These alignments were achieved using a multiple alignment algorithm (Asalign) to align various transcripts of a gene and DNA of the cluster, to reveal various mRNA spliced variants from pre-mRNA. Thus alternative splicing patterns from a group of related nucleotide sequences were created. Then, a topological graph is extracted to use graph theories to study alternative splicing patterns. The topological structure is displayed by a Java applet. The database is accessible via http://166.111.30.65/ASMAMDB.html

2. SpliceDB

SpliceDB is a database of mammalian splice sites. Burset et al has analyzed several characteristics of EST-verified splice sites and build weight matrices for major groups, which can be incorporated into gene prediction programs [5]. They also presented a set of EST-verified canonical splice sites and a set of 290 EST supported non-canonical splice

sites significant for future investigations of the splicing mechanism. The information is accessible via http://www.softberry.com/spldb/SpliceDB.html.

3. PALS

The PALS database took the longest mRNA sequences in 19,936 human Unigene clusters as the reference sequence, and aligned them with EST and mRNA sequences in the same cluster to detect alternative splicing sites [21]. PALs db only marks those alternative splicing sites that have at least 95% identity and 50 bp matches on both ends of an alternative splicing site in an EST-mRNA alignment. Mouse and human EST entries were used in the analysis because these two species have more EST sequences and known mRNA sequences than other databases. The database is accessible via http://palsdb.ym.edu.tw/

4. HASDB

HASDB identifies spliced variants in human genes, through a genome-wide analysis of expressed sequence tags. Human mRNA and UniGene's EST sequences were mapped onto the draft human genome sequence [37]. Splicing is detected by a computational procedure that analyzes the genomic–EST–mRNA multiple sequence alignment. The gene structure is marked on the genomic sequence, based on its alignment with EST and mRNA, by drawing a connection between each pair of genomic letters aligned to a pair of letters in an expressed sequence that are adjacent (i.e. nucleotide i and i+1). Thus, an exon is identified by a contiguous segment of connected letters, an intron by a contiguous segment of unconnected letters and a splice by a connection that jumps from one genomic letter to a distant genomic letter. Thus, a candidate spliced variant is detected as a gap

between two exons that match a single contiguous region of one or more ESTs. HASDB reports spliced variants only for connections that skip >10 bp in the genomic sequence (representing an effective minimum intron length) to screen sequencing errors or alignment heterogeneity artifacts. HASDB predicts alternative spliced variant by detecting them as large inserts in EST data from the publicly available dbEST and UNIGENE. The database, HASDB is accessible via the following URL http://www.bioinformatics.ucla.edu/~splice/HASDB/

5. STACK

STACK addresses the issues of tissue-specific transcripts by merging tissue-specific EST, to provide putative-tissue-specific transcripts for each gene [20]. This in silico method was used to identify exon skipping using verified genome sequence and to study 545 protein-confirmed exon skipping genes on chromosome 22.

6. TAP

TAP is an EST-based gene finder that infers the predominant and alternative gene structures in anonymous genomic sequences [25]. The genomic sequence is searched against dbEST using WU2BLASTN. High-scoring EST hits are aligned to the genomic sequence using SIM4 [18]. Based on genomic EST alignments, TAP predicts predominant gene structure on both strands of the genomic sequence, by detecting the poly-A sites and gene boundaries. The program collected splicing information for 1124 RefSeq human genes. It is accessible via http://sapiens.wustl.edu/~zkan/TAP/

7. ASAP

ASAP, Alternative Splicing Annotation Project is currently the largest human alternative splicing database available, and is expanding to other genomes [30]. It provides information on exon-intron structures of genes, alternative splicing, tissue specificity, and protein isoform sequences resulting from alternative splicing. It has predicted 30,793 alternative splice relationships in human, based on detailed alignment of expressed sequences onto the genomic sequence. ASAP also provides protein isoform sequences for each splice form for experimental validation. ASAP is available at http://www.bioinformatics.ucla.edu/ASAP.

8. ProSplicer

ProSplicer, is a database of putative alternative splicing information derived from the alignment of proteins from Swissprot, mRNA sequences and expressed sequenced tags (ESTs) from Unigene and dbEST against human genomic DNA sequences from ENSEMBL [22]. The database is keyword searchable via gene symbol and provides related reference links to other biological databases and related sequences. ProSplicer is available at http://bioinfo.csie.ncu.edu.tw/ProSplicer/

9. EASED

EASED, an extended alternatively spliced EST database which contains alternative splice forms (ASforms) from nine eukaryotic organisms (*Arabidopsis thaliana*, *Bos taurus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Xenopus laevis*) [40]. It has information such as alternative

splice profile, tissue types, developmental stage, disease and classification of splice events available for human genes. The database is searchable by Genbank accession numbers or keywords such as description, gene names and organism. It also provides an alternative splice profile (ASP) which indicates the number of alternatively spliced ESTs (NAE), the number of constitutively spliced ESTs (NCE) and the number of alternative splice sites (NSS) per mRNA. The quality of predicted alternative splice variants can be assessed by the corresponding NAE and NCE to the EST coverage. The splice propensity of a gene is specified by the NSS value. EASED is accessible via http://eased.bioinf.mdc-berlin.de/

10. ASD

ASD is a database of computationally delineated alternative splice events as seen in alignments of EST/cDNA sequences with genome sequences, and a collection of alternatively spliced exons (experimentally determined) from full-text articles [53]. It contains reported splice events from nine different organisms (human, fly, mouse, worm, rat, chicken, cow and zebrafish and plant) and are annotated for various biological features including expression states and cross-species conservation. These data are available from http://www.ebi.ac.uk/asd/.

Appendix B – Gnomon

Gnomon uses a set of heuristics to find the maximal self-consistent set of corresponding transcript and protein alignment data to set the constraints for an Hidden Markov Model(HMM)-based gene prediction. The program predicts the gene structure in genomic DNA sequences in a multi step fashion. It evaluates the coding propensity of the available transcript alignments and determines their most probable coding regions. A single set of non-overlapping transcript alignments with better coding propensity is chosen. Then the best matching proteins for these transcript alignments are aligned back on the genomic DNA sequence.

Gnomon makes the first pass of the prediction using the above transcript and protein alignments as the constraints. For the transcript alignments, the program makes sure that the chosen coding region is a part of a putative mRNA that can be extended on both sides of the predicted coding region. For the protein alignments, Gnomon checks that the predicted gene has every exon in the right frame as suggested by the protein alignment, although in this case, the program is free to choose the splice sites and to introduce other exons between parts of the protein alignment.

The genes that were built using the alignments from the above step are included in the final output. For the rest of the gene models, the best matching proteins are found and then aligned back on the genomic DNA sequence. These protein alignments are used in the second pass of the prediction for refining the models.

While doing the alignment of the best matching proteins, Gnomon finds all cases where two exons of the protein alignment are within 50 bp and have different frames. Because the probability of such a short intron is extremely low, in all these cases the program introduces a frame shift in the genomic sequence allowing for combining the exons into a single one. In some cases, protein alignments include a stop codon in the middle of the alignment. These stop codons are disregarded during the prediction and appear as premature stops in the model. Both the models with frame shifts and the models with premature stops are annotated as possible pseudogenes in the Gnomon output.