ABSTRACT

Title of dissection: BIOINFORMATIC ANALYSIS OF THE FUNCTIONAL
AND STRUCTURAL IMPLICATIONS
OF ALTERNATIVE SPLICING.

Eugene Melamud, Doctor of Philosophy, 2007

Dissertation directed by: Professor, John Moult,
Center for Advanced Research in Biotechnology,
University of Maryland

In higher Eukaryotes, upon transcription of a gene, a complex set of reactions take place to remove fragments of a sequence (introns) from transcribed RNA. A large macro-molecular machine (the spliceosome) recognizes the ends of introns, brings ends into close proximity and catalyzes the splicing reaction. The selection of the location of the ends of introns (splice sites) determines the final message produced at the end of the process. In some cases, an alternative set of splice sites are chosen, and as a consequence different message is produced. This phenomenon is known as alternative splicing. It is now realized that nearly every Human gene undergoes alternative splicing, producing large variability in types and number of transcripts produced. In this thesis, we examine the functional and structural consequences of alternative splicing on proteins, we look into the mechanism of formation of complex splicing patterns, and examine the role of noise in the process.

BIOINFORMATIC ANALYSIS OF THE FUNCTIONAL AND STRUCTURAL IMPLICATIONS OF ALTERNATIVE SPLICING

by

Eugene Melamud

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Professor John Moult, Chair/Advisor
Professor Steven Mount
Professor Victor Muñoz
Professor Zvi Kelman
Professor Richard Stewart
Professor Steven Salzberg

Table of Contents

List of Figures

Chapter 1

Chapter 1. Introduction

In higher eukaryotes, such as human and mouse, messages transcribed from genomic DNA (pre-messenger RNA) contain large non-coding regions that must be removed before final messenger RNA (mRNA) transcript can leave the cell [1]. An average Human cell contains as many as 800,000 [2] processed mRNAs, nearly every one of which has gone through this intron removal process.

The splicing out of introns occurs cotranscriptionally. That is, as the RNA Polymerase slides along DNA producing new pre-RNA transcript, the introns are recognized and targeted for removal [3]. The removal of introns is catalyzed by a spliceosome, a large macromolecular complex of over 300 different proteins assembled around five small nuclear RNAs (snRNAs). The spliceosome assembles itself at ends of introns, brings the ends into close proximity to each other, and catalyzes two reactions which remove the 5' and then the 3' ends of an intron (for a comprehensive review see Jurica et al. [4]).

It is clear that the choice of the spliceosome assembly sites control selection of exons that will be included in the final mRNA product, but the details and regulation of this mechanism are not well understood. The sequence motifs around the intron-exon boundaries (splice sites) do not in themselves contain enough information to determine transcript structure. A number of algorithms have been developed to predict the location of exons, but these algorithms tend to perform poorly without inclusion of experimental transcript data [5]. One reason for inaccuracy is that the choice of splice sites is influenced by splicing factors, proteins that bind to pre-mRNA in proximity to splice junctions.

These splicing factors typically contain one or more RNA binding domains and interact directly with short regulatory sequence elements located in both exons and introns [6]. Some of these factors prevent spliceosomes from recognizing splice sites, others enhance recognition. One large class of splicing factors is SR proteins, (serine- and arginine-rich proteins). These proteins can play both positive and negative roles in selection of splicing, depending on the location of their binding [7]. Another large class of splicing factors is heterogeneous nuclear ribonucleoproteins (hnRNP) proteins. These proteins typically bind to regulatory regions within introns and act to inhibit recognition of splice sites [8]. In the end, it is the balance of concentration and activity level of enhancers and silencers that determines how often specific splice sites will be chosen [9].

There are three possible outcomes of alternative splice site selection. First, a different 5' end of an intron might be chosen; second, a different 3' end of an intron might be chosen, and lastly, an intron may not be recognized at all - causing an intron retention event. Changes to introns result in changes to exon structure such as: exon indels, or exon 5' or 3' end modification. Differences in final mRNA sequences may be classified into three categories: changes before the translation initiation site (5' UTR), changes after the translation termination site (3' UTR), and changes inside protein coding regions. The majority ($\sim 70\%$) of alternative splicing events affect protein coding regions[10]. Thus a large number of alternative protein products can potentially be generated.

One well known example of generating of protein diversity through alternative splicing is the Drosophila Dscam gene [11], which has a primary function in axon guidance. It is estimated that 38,016 different protein molecules could be generated from this gene and evidence suggests that many are expressed and are functionally important [12]. Similarly, another group of brain expressed genes, Neurexins, which function as receptors of neuropeptides, are estimated to generate on the order of 1000 different isoforms [13]. Another cell surface receptor protein, CD44, with diverse functions such as cell-cell

recognition and cell signaling, can potentially produced 1024 different combinations, of which 69 have been detected [14].

Alternative splicing need not directly express a large diversity of proteins to be functionally important. A famous example of the effect of alternative splicing on phenotype is the Sex-lethal (Sxl) gene in Drosophila. This gene is a splicing factor, that acts as a master switch to activate production of a functional protein product of transformer (tra) gene, which in turn activates a pathway resulting in female differentiation [15]. Another set of splicing factors, Nova1 and Nova2 (mouse), are tissue specific factors that regulate a large network of brain specific targets essential for neuron viability [16, 17].

Until fairly recently, the extent of alternative splicing in higher eukareotes has been under-appreciated. The availability of a Human genome draft sequence and rapid growth in the number of expressed sequence tag (EST) libraries has led to the discovery that alternative splicing affects 35 to 40% of all Human genes [18, 19, 20]. More recent estimates based on microarray experiments put the value at a minimum of 74% [21]. The true extent of alternative splicing remains unknown, as only a small fraction of all transcripts have been sampled by EST experiments, and only a limited set of exon-exon junctions are present in microarray platforms.

It has been suggested that one explanation for the existence of a large number of alternative isoforms is that alternative splicing provides complex organisms such as Human with a mechanism for generating functional diversity from a relatively small set of genes [22, 23]. Indeed, compared to yeast, which has only a few alternatively spliced genes, the frequency of splicing in higher eukaryotes is significantly higher [24]. It would seem that alternative splicing could potentially be responsible for an increase in complexity, but before accepting this hypothesis, it must be shown that alternative splicing provides novel functions.

Many bioinformatics studies have looked at various aspects of alternative isoforms. Although designs and datasets vary, the general conclusions are similar: At least 2/3 of all Human genes are alternatively spliced. A small fraction (10-20%) of all alternative splicing events are conserved across multiple species [25, 26, 27]. Approximately 20-40% of alternative splicing events show tissue specificity [28, 29]. Approximately 70% of changes affect protein coding regions [10]. Conserved alternative splicing events have different characteristics from species specific splicing events, such as: they are expressed in greater abundance, have a tendency to preserve coding frames, and have lower synonymous substitution rates[30, 31, 32]. A large number (at least 35%) of alternative isoforms result in truncated proteins and are thus subject to degradation by nonsense mediated decay (NMD) [33].

Microarray based analyses of alternative splicing are largely in agreement with cDNA/EST analysis. Johnson et al. using exon-junction arrays showed that at minimum 74% of genes are alternatively spliced [21]. Pan et al. [34] looked at tissue specific expression in mouse of alternative exons and found that most ($> 70\%$) splicing events are not tissue specific and expressed at low abundance. In contrast to EST/cDNA based predictions, analysis of NMD events via microarray experiments showed that there are steady levels of alternative transcript with premature stop codons and few seemed to be regulated by NMD [35].

Although there seems to be a consensus in the field that conserved alternative splicing events are probably functional, there is no agreement on functionality of species specific isoforms. It has been suggested that minor isoforms need not provide functional components, but may act to regulate abundance levels of major isoforms [36]. It has also been suggested that the functional role of alternative splicing can not be understood on a per gene basis, changing concentration of various splicing factors results in global splicing changes, and the functional implications at the system level effect remain hidden

[37]. Of course, without a complete picture of all interactions in the cell it is impossible to evaluate the truth of this hypothesis.

An alternative explanation for the large number of alternative isoforms is that they are produced as consequence of noise in the splicing machinery. [23, 38, 39]. Assuming that splicing machinery is not perfect and makes occasional mistakes in selection of splice sites, a large number of isoforms could be generated by the accumulation of errors. Most of these isoforms will not be functional, but as long as they are not toxic and adequate levels of normal gene product are produced, there will be little selection pressure to remove them. In chapter 3 of this thesis we show that the observed isoform diversity can be reproduced with a simple error model. We find that there is pressure to reduce the frequency of errors in highly expressed genes and in genes with many introns. In Chapter 4, we examine the impact of alternative splicing on protein structure. In agreement with the noise hypothesis, we find that species specific isoforms usually result in unstable fold conformations. In a slight deviation from the above topics, in Chapter 2, we also examine the role of interaction between the splicing and transcription machinery in the formation of alternative splicing patterns.

Chapter 2

Chapter 2. Alternative Splicing Patterns

2.1 Abstract

The process of formation of a final mRNA transcript involves a set of complex interactions between spliceosomal and transcriptional components. The types and frequency of alternative splicing and alternative transcriptional events are determined by these interactions, but as yet they are poorly understood. Furthermore, many transcripts are the outcome of a combination of multiple alternative events, and it is unclear if these events are coordinated, or if so, how. To help address these questions, we have developed a symbolic language for describing differences between transcripts. We have analyzed splicing patterns in Human genes and identified statistically significant correlated splicing and transcription events. The primary findings are as follows: (1) Most splicing patterns can be explained by independent selection of splice sites within a single spliceosome; (2) There is little coordination between adjacent spliceosomal complexes in formation of splicing patterns; (3) Alternative transcription events are the dominant source of transcript diversity; (4) There is strong linkage between alternative transcription and alternative splicing events.

2.2 Introduction

In Eukaryotes, there are three major mechanisms that contribute to the diversity of transcripts: alternative splicing, alternative transcription initiation, and alternative tran-

scription termination [40, 41]. Alternative splicing is a consequence of a spliceosome operating on an alternative set of splice junctions during the process of intron removal [6]. Alternative 5' end formation is a result of alternative promoter selection by the RNA polymerase complex [42]. Alternative 3' end formation is a result of selection of alternative poly(A) sites by spliceosomal and termination/cleavage machinery [43].

The traditional view was that both 3' end formation and splicing are post transcriptional events, but it is now clear that these processes occur co-transcriptionally [44, 45, 46, 3]. The final mRNA transcript is thus the outcome of interaction between a number of concurrent processes that involve a multitude of components from the transcription and splicing regulatory pathways. Interactions between components can be complex, with alternative promoters causing changes in splicing [47], polymerase processivity causing alternative splicing [48], splicing factors enhancing polyadenylation [49, 50] and the presence of proximal 5' splice sites enhancing transcriptional initiation [51]. (Illustrated in Figure 2.1)

The mechanism of action of a single spliceosome has been worked out in great detail [6]. However, formation of large splicing patterns that involve many introns remains poorly understood. In particular, it is unknown whether there is significant communication between multiple spliceosome assemblies in the formation of splicing patterns involving multiple introns. Furthermore, very little is currently known about the strength and frequency of interaction between transcription and splicing machinery. How strong are these interactions and how frequently do the properties of one influence the behavior of the other?

To address these questions, we have developed a symbolic representation of differences between isoforms. The formalism is an extension of that typically used to describe alternative splicing events, introducing more detailed categories and adding a symbolism for alternative transcription events. We have compiled statistics on splicing patterns in

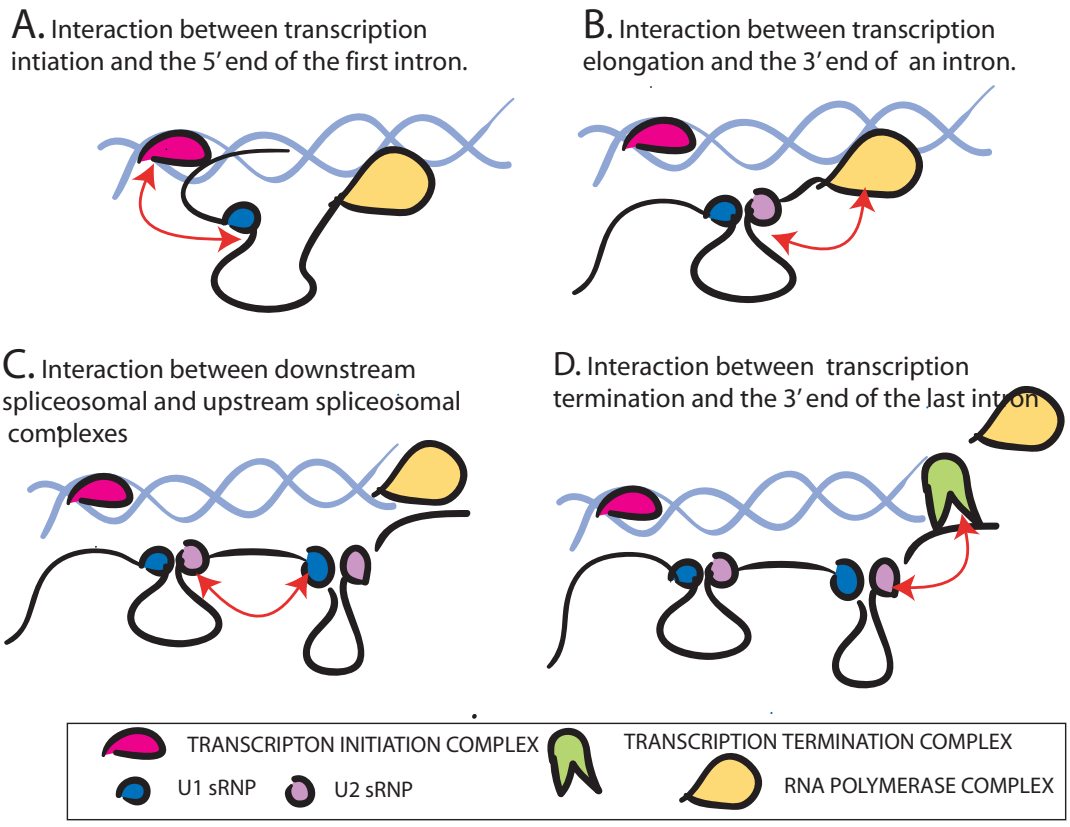

Figure 2.1: Examples of possible interactions between splicing and transcription machinery. (A) Interaction between transcription initiation and a spliceosome. (B) Location of 3' intron end determined by location of RNA Polymerase. (C) Exon definition mechanism: U1 and U2 snRNP communicate across exon. (D) Interaction between a spliceosome and 3' end formation.

Human transcripts and developed a model that simulates the formation of random exon patterns. The random model allows us to derive accurate statistics on expected frequencies and correlations between various transcription and splicing events.

Our analysis indicates that most splicing patterns are formed by independent selection of splice sites within a single spliceosomal assembly. We do not find evidence for coordination between adjacent spliceosomal assemblies. However, correlated patterns formed by transcription initiation/termination and spliceosomes are highly enriched.

2.3 Results

2.3.1 Generation of Splicing Patterns

We describe differences between isoforms using a symbolic character representation of exon structure. Our aim is to produce a formal language description of alternative splicing events similar in form to the amino acid code. For example, if two isoforms contain five identical exons, we write down a string of length five: 'IIIII', where each 'I' character represents an identical exon. If a minor isoform of this gene skips an internal third exon, the corresponding pattern is 'IIAII', where 'A' represents an exon deletion relative to the major isoform.

In all, we define an alphabet of fourteen different symbols. Six upper case symbols ('I','A','B','5','3','X') represent differences that are exclusively due to changes in exon structure introduced by alternative splicing. Six lower case symbols ('a','b','o','e','y','z') represent differences due to a combination of alternative transcription initiation/termination and alternative splicing events. In addition to symbols used to indicate differences between exons, we include two symbols 'S' and '*' to mark the beginning and ends of patterns. The alphabet is illustrated in the Figure 2.2.

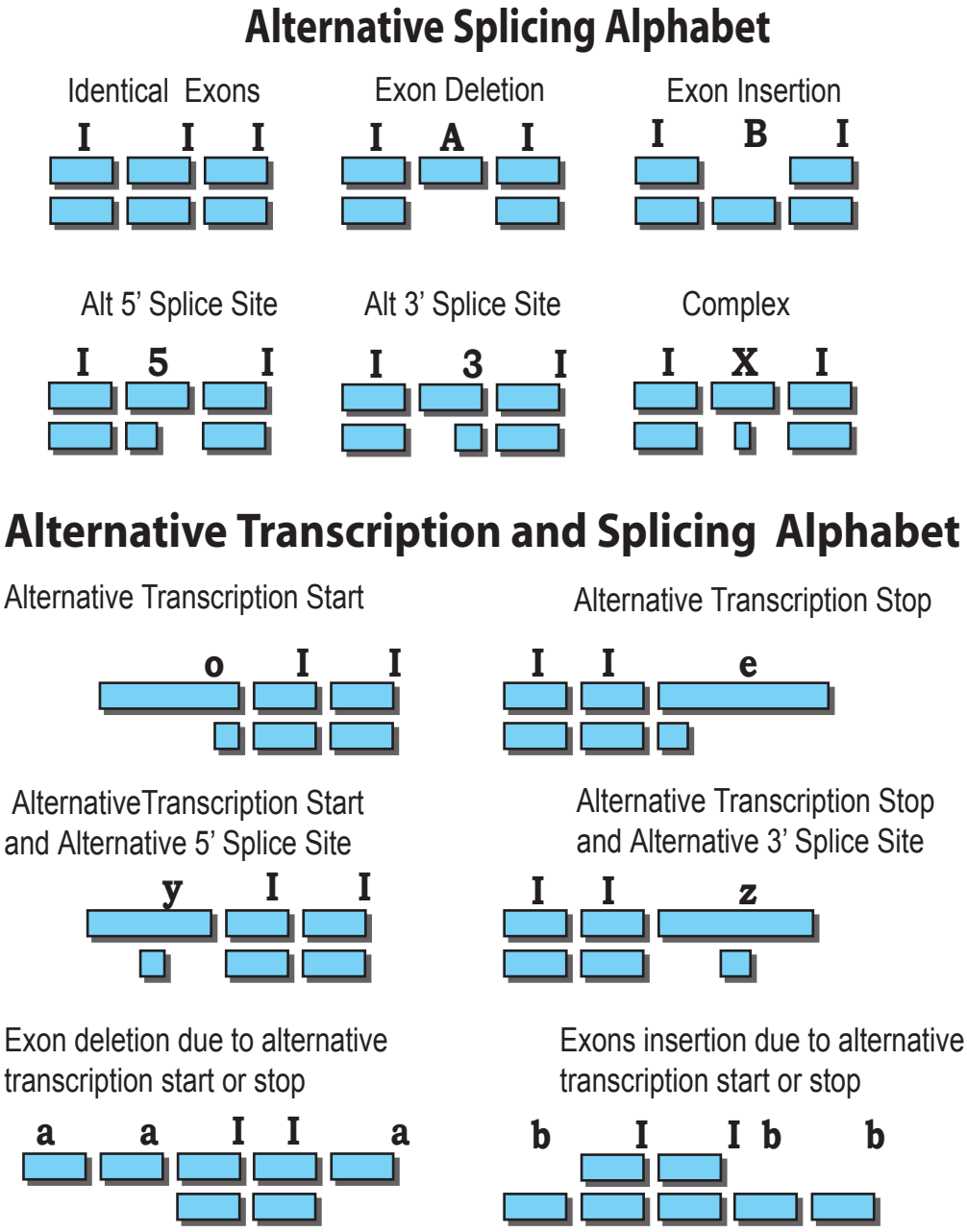

Figure 2.2: The alphabet used to describe differences between major and minor isoforms. The symbols are broken down into two groups: differences that are due to selection of an alternative splice site by splicing machinery (symbols 'A', 'B', '5','3', 'X'), and differences that are due to splicing machinery and/or transcription machinery (symbols 'a', 'b', 'o', 'e', 'y', 'z'). In each case, the top row shows a major isoform exon/intron structure, and the bottom row shows the structure of a related minor isoform.

The algorithm for symbol assignment is described below. We identify equivalent pairs of exons between the major isoform of each gene and each minor isoform, using the genomic coordinates of the exons. Specifically, two exons are considered equivalent if their alignments to the genome sequence overlap. All exons within an alignment are classified into two categories: exons that differ solely as a result of change of splice site choice by spliceosomes (internal exons), and exons that differ as a result of changes introduced by transcription initiation/termination machinery or due to a combination of splicing and transcription initiation/termination machinery (external exons). Illustrative examples are shown in Figure 2.3.

Internal exons (those with genomic co-ordinates at least partly overlapping with the equivalent exon) are labeled with upper case symbols. External exons are labeled with lower case symbols. Symbols are assigned as follows:

1. Both exons have identical genomic coordinates 'I'
2. One of the exons is missing in the major or the minor isoform
    (a) Missing in the minor isoform - exon skip
        i. Missing internal exon - upper case 'A'
        ii. Missing external exon - lower case 'a'
    (a) Missing in the major isoform - exon insertion
        i. Inserted internal exon - upper case 'B'
        ii. Inserted external exon - lower case 'b'
3. Difference in 5' boundary between equivalent exons
    (a) Internal exon, alternative 5' splice site - symbol '5'
    (b) Alternative transcription start site - 'o'
4. Difference in 3' boundary between equivalent exons
    (a) Internal exon - alternative 3' splice site - symbol '3'
    (b) Alternative transcription termination - 'e'
5. Difference at both the 3' and 5' ends of equivalent exons

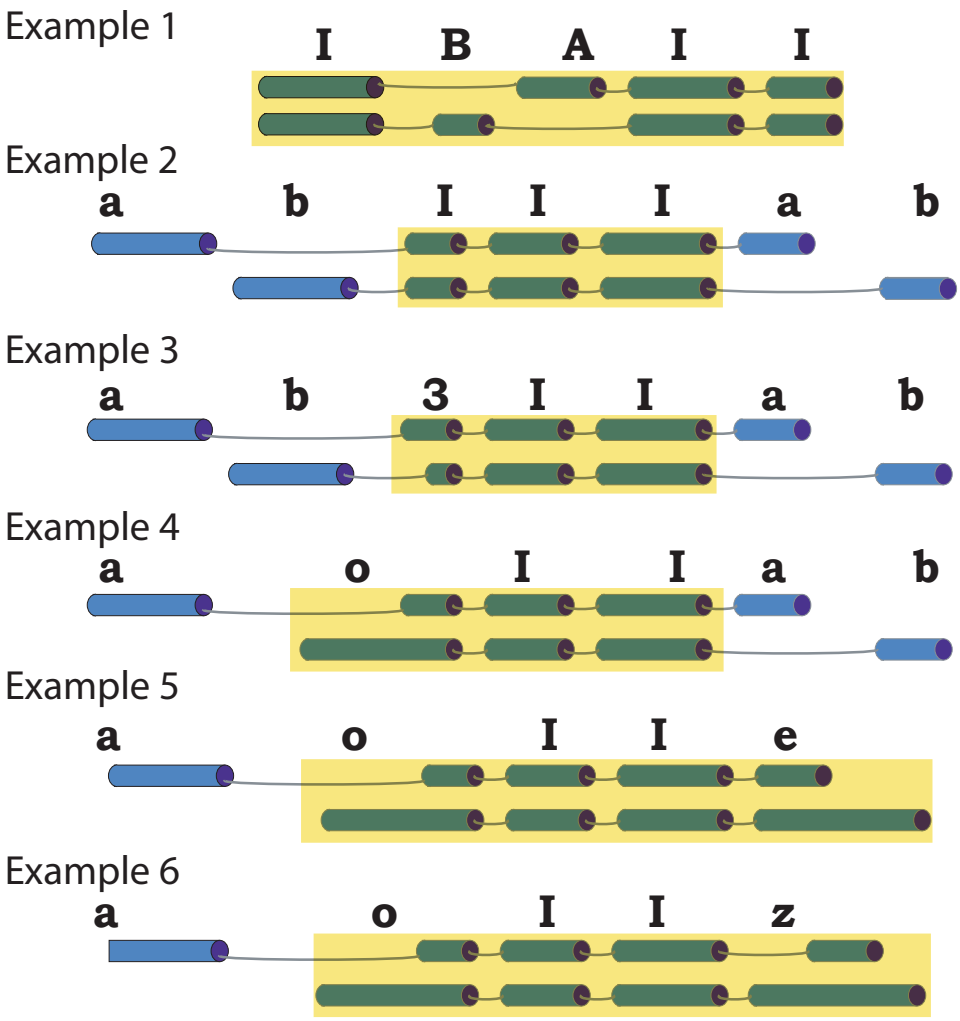

Figure 2.3: Illustrative examples of symbol assignment. In each case, the top row shows the exon structure of a major isoform, and the bottom row the exon structure of a minor isoform of the same gene. Exons are divided into two categories, internal exons (highlighted yellow) and external exons. All exons to the left of the first overlapping exon pair and all exons to the right of the last overlapping pair are considered external. Differences within internal exons are largely due to alternative spliceosomal events, while differences in external exons are due to interaction between spliceosomal and transcription machinery. Example 1 shows an internal exon swap, encoded as the string 'IBAII', generated by spliceosomal machinery, while the two exon swaps in the second example 'abIIIab' are generated by both transcription and splicing machinery. Selection of an alternative 3' splice site by a spliceosome generates the 'ab3' pattern in Example 3, while alternative transcription initiation of the same exon generates the 'ao' pattern (Example 4). Examples 5 and 6 illustrate similar modifications of internal exons due to alternative termination.

(a) Internal exon - symbol 'X'
(b) Alternative transcription start and 3' alternative splicing - 'y'
(c) Alternative transcription termination and 5' alternative splicing - 'z'

There are three points to be noted. First, although we only use full length cDNA sequences, the ends of transcripts might still be poorly defined. To avoid small differences at the ends of transcripts inflating the 'o' and 'e' counts, we require a difference of at least 20 nucleotides for these symbols to be assigned. Second, the alphabet describes all possible differences in exon structure, with the exception of intron retention events (events where an exon-intron-exon in one isoform becomes a single continuous exon in another). Although some intron retention events might be biologically significant, we cannot distinguish them from experimental artifacts such as incomplete processing of pre-mRNA. Third, the symbols '5' and '3' refer to the ends of introns, rather than the ends of exons (a possible alternative choice), reflecting the units the machinery operates on.

Examples of splicing patterns in the caspase family are shown in Figure 2.4.

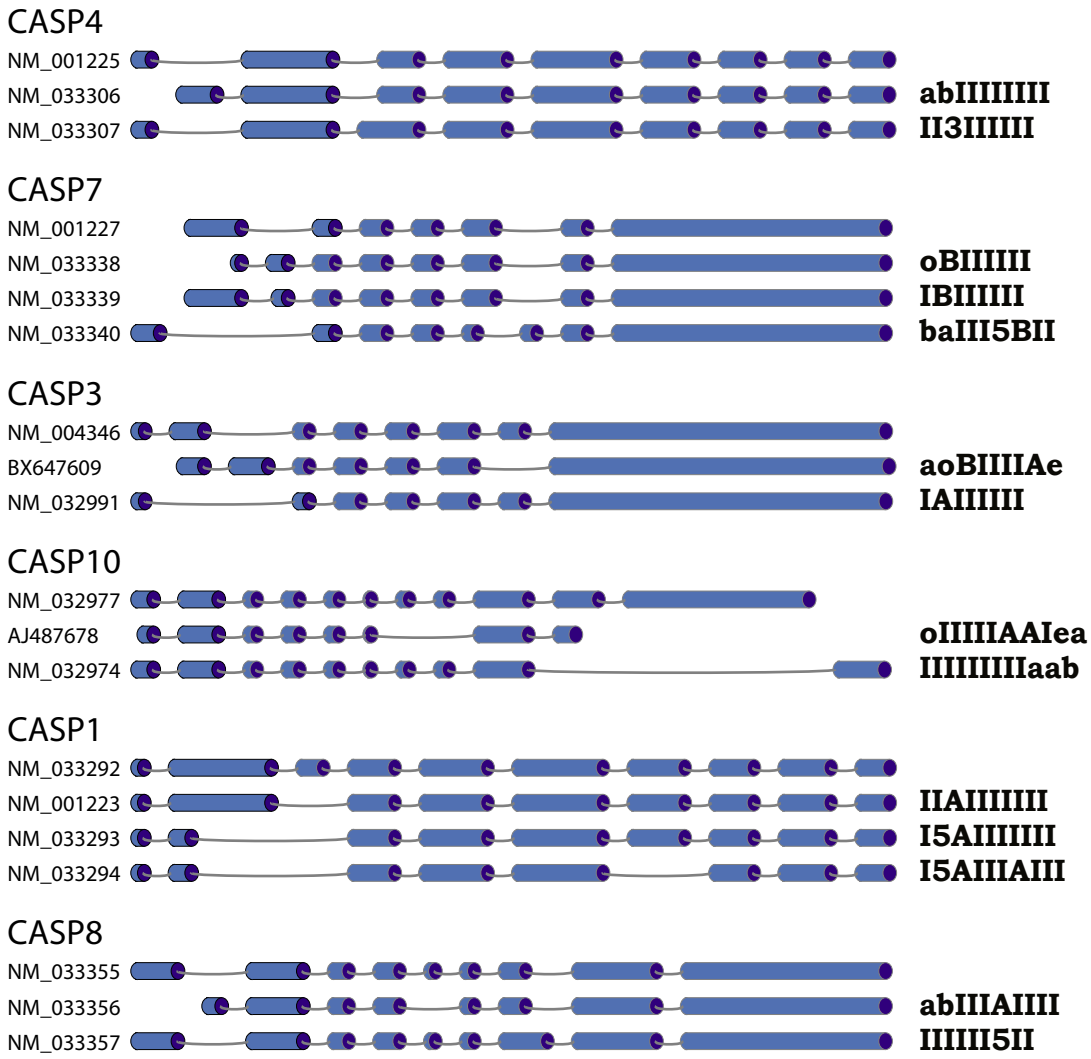

Figure 2.4: Examples of strings describing alternative isoforms in Human Caspases. The second isoform of CASP4 contains a novel exon generated by an alternative transcription initiation event, resulting in the formation of an exon swap pattern ('ab') at the beginning of the gene. The exon swap pattern due to alternative promoters also occurs in CASP7 ('ba') and in CASP8 ('ab'). Alternative termination events in CASP10 result in two exons being replaced by a single one, forming an 'aab' pattern. There are also many examples of internal exon insertions and exon deletions in the caspases, forming patterns with 'B' and 'A' symbols. For example, the second isoform of CASP10 is missing two internal exons relative to the major isoform, generating an 'AA' pattern. Exon insertions and deletions are frequently associated with 5' splice modification, forming '5A' and '5B' patterns, as seen in CASP1 and CASP7. The alternative transcription initiation event in the second isoform of CASP3 produces an 'ao' pattern, while an alternative transcription termination event in the second exon from the end in CASP10 produces an 'ea' pattern.

2.3.2 Overview of Splicing Patterns

A reliable non-redundant set of isoforms was generated, starting from an initial set of 22,000 genes and 136,000 cDNA transcripts. Some genes were removed because they had only a single cDNA transcript or because all transcripts contained errors. For the remaining genes, one of the transcripts was selected as the major isoform. After filtering out duplicate transcripts and partial transcripts, we were left with 29,543 potential minor isoforms. For the analysis, we require a set of unique splicing events, rather than unique full transcripts, and the set of unique minor isoforms of a gene may contain multiple instances of the same event. For example, there may be two isoforms with different 5' ends, both containing the same exon skip in the middle of the transcripts. To ensure each event is only counted once, we randomly selected only a single minor isoform for every gene, resulting in a final set of 10,305 patterns. These patterns were used to compute the frequency of each symbol (Table 2.1) and of pairs of symbols (Table 2.2).

Some care is necessary in interpreting the observed fractions of each symbol type. Symbols 'S' and '*' are included to indicate start and stop of patterns, and do not represent changes to the exon structure of a gene, but do contribute to the total number of observed symbols. It is also important to point out that observed fractions do not represent actual probabilities of various exon types, because we are only looking at splicing events within our dataset of 10,305 minor isoforms and not in the whole transcript space. If all transcripts from all isoforms were considered, the number of exons that are different between transcripts would shrink, while the frequency of identical exons would increase to approximately 95% of all exons. In the context of this application, we compare exon frequencies to each other, and thus it is relative frequencies of the symbols that are important, not their absolute values.

Symbol	Count	Frequency	Description
X	3	0.002%	Complex (both sides of internal exon differ)
z	56	0.04%	Complex (both sides of the last exon differ)
y	160	0.13%	Complex (both sides of first exon differ)
5	472	0.37%	Alternative 5' splice site
B	831	0.65%	Alternative 3' splice site
3	860	0.67%	Exon insertion (internal exon)
b	1494	1.17%	Exon insertion (external exon)
A	1977	1.55%	Exon deletion (internal exon)
e	3924	3.08%	Alternative transcription termination
a	5384	4.22%	Exon deletion (external exon)
o	6412	5.03%	Alternative transcription initiation
$\star$	10305	8.08%	End of pattern indicator
S	10305	8.08%	Start of the pattern indicator
I	85285	66.91%	Exon identical in both isoforms
	127468	100%	

Table 2.1: Exon change (single symbol) frequency. Frequency is defined as the ratio of symbol to total symbol count. Note symbol 'S' and '*' indicate pattern start and end - not exon differences.

As can be seen in Table 2.1, only 4% of all changes to exon structure are solely due to spliceosomal events ('A','B','5','3','X'). Excluding, 'S', '*' and 'I' symbols, all other differences are due to alternative transcription or a combination of alternative transcription and splicing. These combined events ('a','b','o','e','z','y') represent 20% of all exon changes. It is important to point out that our estimates of the number of exons modified due to alternative transcription initiation and termination are conservative. To avoid contamination of the dataset by partial cDNA sequences, we have removed transcripts that are missing more than 30% of the exon structure relative to the major isoform. Since some genuine shorter isoforms were affected by the filter, we underestimate counts for 'a' and 'b' symbols.

Alternative transcription termination events resulting in changes at the 3' end of the last exon ('e*') or 3' terminal exon swaps ('ab*', 'ba*' 'abb*' etc.) occur in 40% of all patterns (4228 out of 10,305 patterns). This finding is similar to finding by Tian et al.[43] who found 50% of all human cDNAs have alternative 3' ends. Modification of an exon

by an alternative promoter is more common than by alternative transcription termination. Alternative transcription initiation events ('So', 'Sa', 'Sb', and 'Sy') are found in 77% of all patterns (7968 out of 10,305 patterns). This estimate is somewhat larger than the 56% value obtained in cap analysis of Human cDNA (CAGE) experiments [52].

Among spliceosomal induced changes ('A','B','5','3','X'), the most common is exon deletion 'A'. Deletions are observed twice as often as insertions ('B'). One sided modification of an exon through an alternative 5' splice site or an alternative 3' splice site are the next most common events. Exons modified on both 5' and 3' ends ('X') were observed only 3 times.

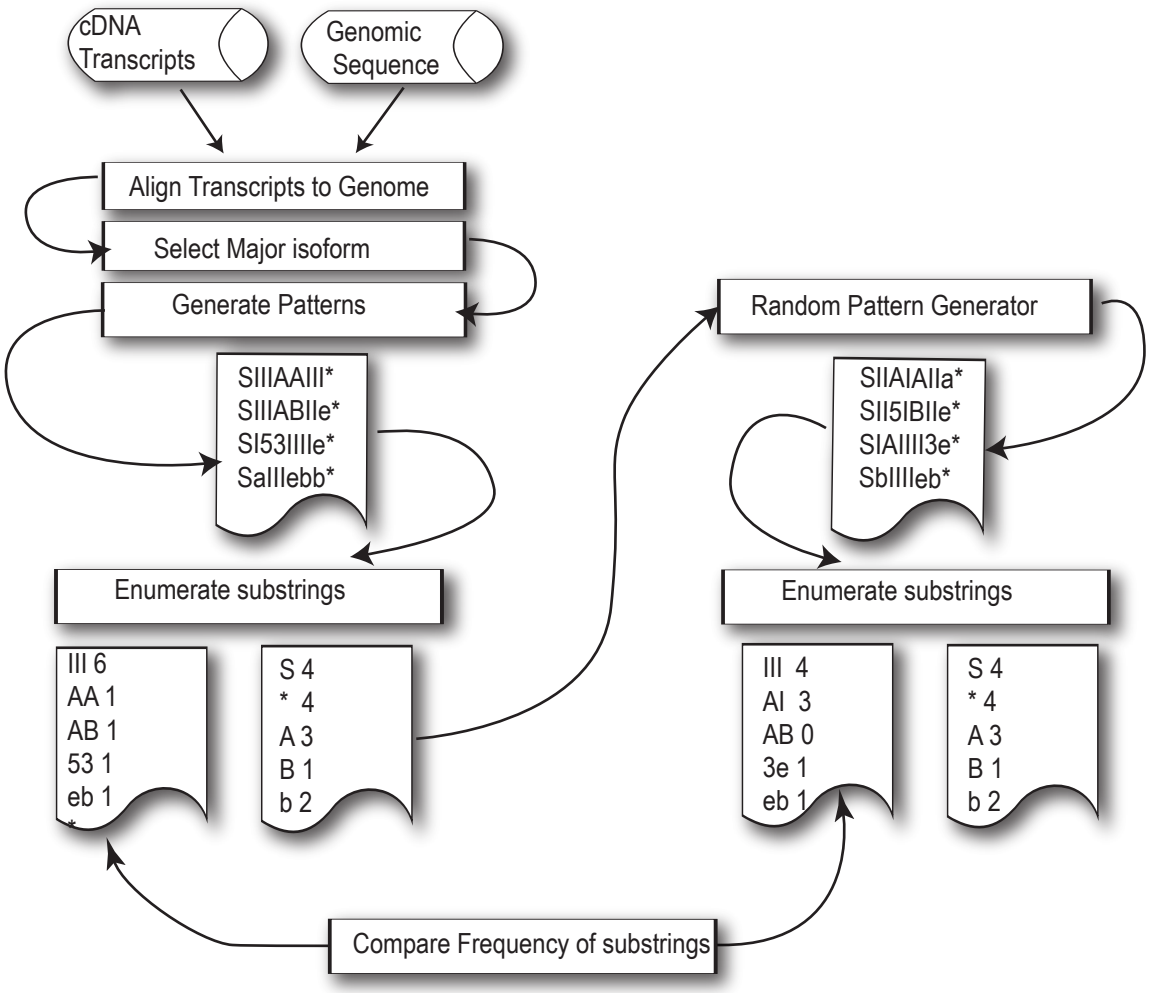

Figure 2.5: Overview of steps used to compute and compare real and random patterns.

2.3.3 Spliceosomal Patterns

We next compare the observed frequency of multi-symbol patterns with that expected by chance. There are a number of position and sequence restrictions on possible patterns that affect the expected random frequency of symbol pairs. For example, external exon deletions (symbol 'a') can only occur at a subset of positions on the ends of transcripts, and such an external deletion can never be followed by an internal deletion symbol 'A'. To reliably calculate the random frequencies we use a procedure that simulates random patterns, taking into account positional biases. The algorithm generates random patterns of the same length and composition as the observed patterns. Random patterns are checked against a set of rules and only the subset of patterns that pass are kept. We then compile statistics on the frequency of random substrings and compare them to observed frequency of substrings in real patterns. (Procedure illustrated in Figure 2.5)

The comparison between observed and random substring frequencies is show in Table 2.2. Double exon skips (AA) and insertions (BB) tend to occur together much more frequently than expected by chance (11-fold and 9-fold enrichment respectively). Longer runs of insertion are less common than longer runs of deletion (Table 2.3), with only 6% of all insertions occurring in multiple 'BB' patterns, compared to 18% of all deletions occurring in multiple 'AA' patterns. Of course, the distinction between insertions and deletions is slightly subjective, depending on the choice of major isoform. As shown in Figure 2.6, both these patterns arise as a consequence of a single spliceosome operating a one different splice site. Exon insertion or deletion events are often coupled to modification of 5' and 3' boundaries of adjacent exons. For example, '5A' , 'A3', '5B' and 'B3' patterns all occur at a frequency higher than expected by chance. As illustrated by Figure 2.6, in all cases, strong association between deletion and exon modification only requires a single spliceosome to operate on one different single splice site.

Pattern	Observed	Expected	Enrichment	Description
AA	463	41+/-6	11.13	Internal double exon skip
BB	67	7+/-2	9.05	Internal double exon insertion
ab	572	71+/-9	8.04	Exon swap at the ends
ba	438	69+/-6	6.33	Exon swap at the ends
S5	120	21+/-3	5.64	Alternative 5' splice site (ss) of the first intron
53	22	4+/-3	4.75	Intron modified at both 5'ss and 3'ss
5A	43	10+/-3	3.94	Alternative 5'ss and deletion
YA	13	3+/-1	3.79	Alternative promoter and deletion
oe	136	35+/-7	3.78	Two exon transcript: Alternative start followed by alternative stop
5B	15	4+/-1	3.17	Alternative 5'ss and insertion
A3	55	20+/-5	2.70	Deletion and 3'ss
aa	2459	1122+/-17	2.19	External double exon skip
B3	19	8+/-2	2.13	Insertions and alternative 3'ss
bb	235	116+/-9	2.02	External double insertion
3*	120	61+/-6	1.94	Alternative 3'ss in the last intron
bI	524	275+/-11	1.90	Exon insertion at 5'end of transcript
AB	33	18+/-4	1.79	Spliceosomal Exon Swap
BA	26	15+/-4	1.71	Spliceosomal Exon Swap
oB	107	64+/-7	1.65	Alternative initiation and insertion of an exon.
SY	146	90+/-11	1.62	Alternative initiation and alternative 5'ss in the first intron

Table 2.2: Comparison between the observed and expected frequencies for patterns of length two.10,306 complete and randomly generated patterns were broken down into overlapping fragments of length two, and fragments were counted. For example, the full length pattern 'I5AAAI' is broken down into fragments 'I5', '5A', 'AA', 'AA', 'AI'. Expected frequencies were computed as an average over 100 simulations. Only patterns observed at least 10 times are included.

Deletions	Observed	Insertions	Observed
A	1232	B	713
AA	202	BB	43
AAA	42	BBB	3
AAAA	22	BBBB	2
AAAAA	4	BBBBB	3
AAAAAA	2	BBBBBB	0
AAAAAAA	1	BBBBBBB	0
AAAAAAAA+	1	BBBBBBBB+	0
TOTAL	1514	TOTAL	764

Table 2.3: Frequency of multiple deletions and insertions in the observed data. Longer runs of deletion (more than two consecutive 'A's) are more common than longer runs of insertion (more than two consecutive 'B's). Approximately 18% of all 'A's occur in multiple runs, in contrast to 6% of all 'B's, suggesting that insertions are more independent events.

One of the most interesting enriched patterns is alternative splicing that modifies both sides of an intron producing a '53' pattern, a longer ones such as '5A3', '5B3' and '5AA3'. Although these events are rare, they tend to occur at frequencies higher than expected by chance (enrichment > 5). How these patterns arise is not clear. The recognition of the 5' and 3' ends of introns are two independent events [6], and interaction between the U1 and U2 snRNPs takes place only after the splice sites have been recognized. However, enrichment of these patterns indicates that a change in a 5' splice site is sometimes coordinated with a change in the neighboring 3' splice site. It is possible that '53' patterns are generated through two intron removal reactions. For example, where two introns overlap, with one intron inside another, as illustrated in Figure 2.8. The limitation of the intron inside intron model is that it can only explain introns that are nested, however approximately 40% of all '53' patterns in our dataset were found in a staggered configuration.

It is also possible that '53' patterns are generated by U12 spliceosomes. The U12 spliceosome is capable of recognizing 'AG-GT' type introns and has been shown to bind

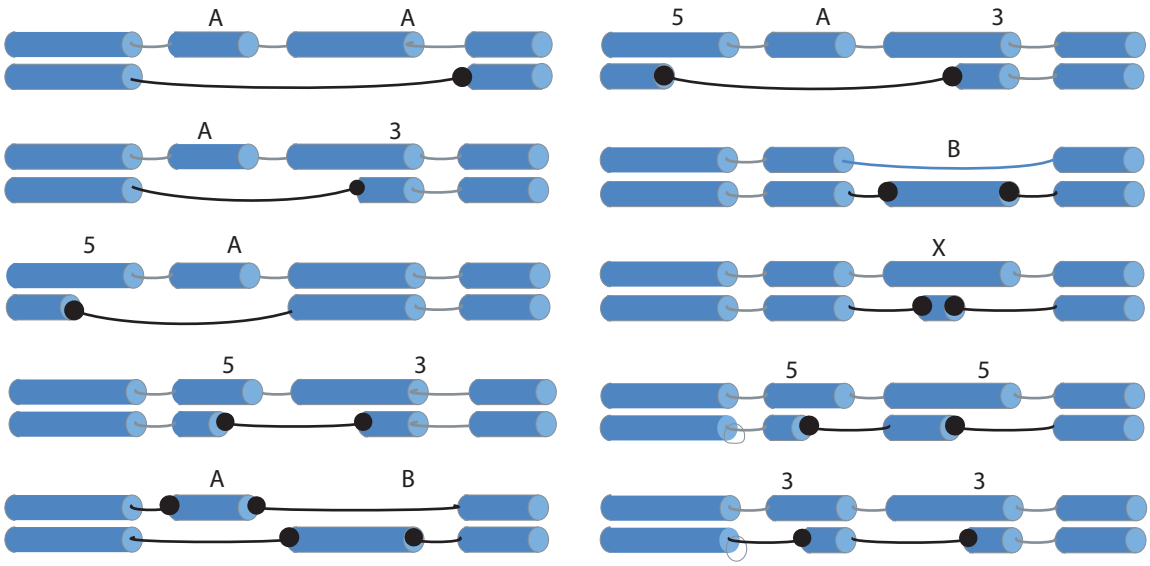

Figure 2.6: Patterns generated by spliceosome machinery. Splice site changes are indicated by circles. The most frequently observed patterns, such as 'AA', 'A3' and '5A', can all be generated by change in a single splice site choice by a single spliceosome, and occur at frequencies higher than expected by chance. The '53' and '5A3' patterns are generated by a single spliceosome complex, but modify both the 5' and 3' splice sites of an intron. The '53' pattern is rare, but occurs at a frequency higher than expected by chance. The patterns '55' and '33' involve two adjacent spliceosomes. They are rare and occur at expected random frequencies. The exon swaps 'AB' and 'BA' are complex patterns that involve both deletion and insertion of exons, and therefore require coordination of four splice sites. These exon swaps are rare, but happen at frequencies higher than expected by chance. Exon insertion 'B' requires definition of two splice sites from adjacent spliceosomes.

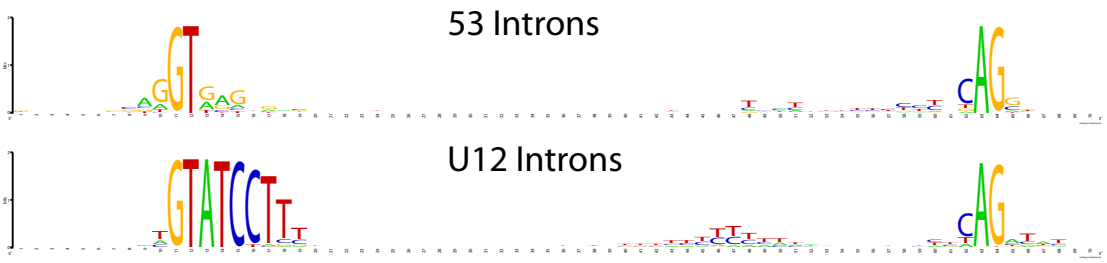

Figure 2.7: Motifs around '53' introns and U12 'AG-GT' introns. U12 introns were obtained from the U12DB database [54]. Sequence logo pictures were constructed using Weblogo program [55, 56].

to both 5' splice site and branch point cooperatively [53]. Figure 2.7 shows the sequence logo comparisons between sequence motifs around '53' introns and U12 'AG-GT' introns, obtained from the U12DB database [54]. It is clear that sequence motifs around '53' are significantly different from U12 sequence motifs, indicating that U12 is not the cause for enrichment of the '53' pattern.

The 'X' pattern arises when both 5' and 3' side of the same exon are modified by adjacent spliceosomes. 'X' exons are extremely rare, with only 3 examples in the whole dataset of 106,855 exons. The random expectation for 'X' patterns is the product of alternative 5' splice probability and alternative 3' splice probability, which is in fact also 3 events in 106,855. It is hard to reconcile this low frequency with the exon definition model [57], where the 5' and 3' ends of an exon are defined by communication between two spliceosome assemblies on either side of the exon. If there is such communication, we would expect to find more 'X' patterns. In addition to 'X', patterns '33', '55' are also produced by alternative splicing events in two adjacent spliceosomal assemblies. Both of these events are extremely rare and observed at the frequency expected by chance: 9 examples for '33', and 3 examples of '55'.

Generation of exon swaps ('AB' and 'BA' patterns) requires coordination between

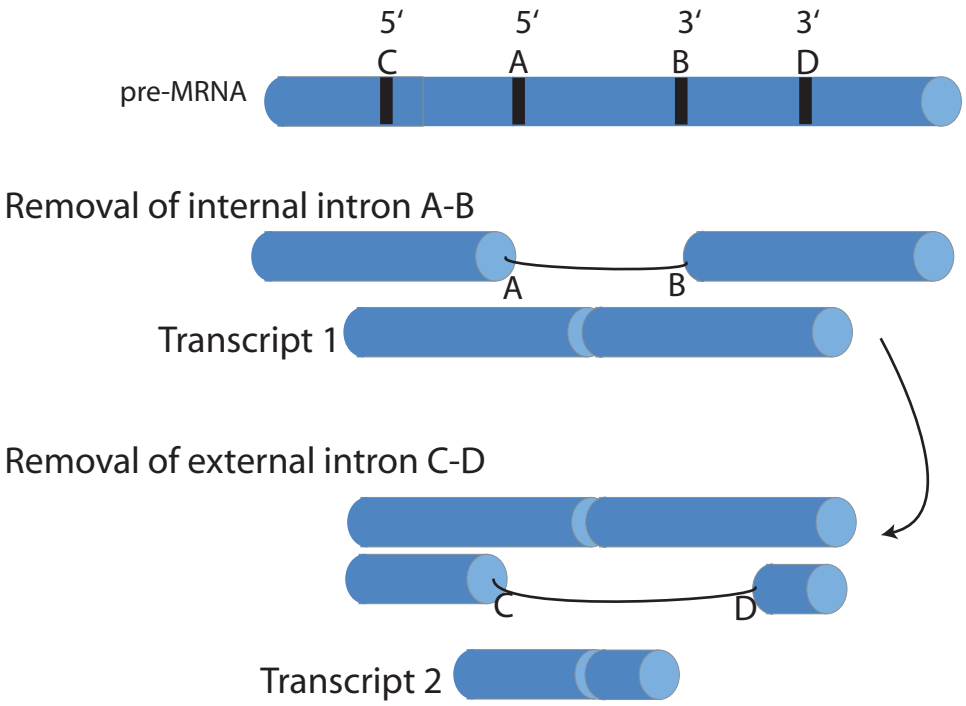

Figure 2.8: Possible mechanism for formation of the '53' pattern. Two overlapping introns 'A-B' and 'C-D' are removed in two splicing reactions. The first splicing reaction, removes intron A-B, forming the first product (transcript 1). Transcript 1 is used as a substrate for a second splicing reaction, which removes the second intron C-D. If both products are detected and compared to genomic sequence, the change in both 5' and 3' splice will be observed.

four splice sites, two of which must be activated in the alternative, and two that must be suppressed. These events are rare (59 examples total), but are observed at higher than expected frequency, by a factor of 1.7. This enrichment indicates that coordinated removal of adjacent introns is possible, although infrequent.

2.3.4 Transcription Machinery Patterns

Multiple deletions ('aa') and multiple insertions ('bb') of exons due to alternative transcription initiation or termination are the most frequently observed types of double modification to exon structure (Table 2.2). Deletions occur more frequently than insertions. Although we see mild (2 fold) enrichment in these patterns, this is simply a reflection of the fact that many exons can be changed with a single change in location of transcription start or end. Long runs of multiple exon deletions and insertions occur at the 5' end much more frequently than at the 3' end (Table 2.4). It is not immediately obvious why transcription initiation tends to delete/insert more exons.

Transcription Initiation	Observed	Transcription Termination	Observed
a	929	a	204
aa	285	aa	55
aaa+	352	aaa+	96
b	151	b	52
bbb	29	bb	9
bbb+	19	bbb+	1
ab or ba	716	ab or ba	105
a+b+ or a+b+	34	a+b+ or a+b+	80

Table 2.4: Comparison between exon deletions and insertions introduced by alternative transcription or termination. ('+' indicates a string length equal or longer)

A mix of insertions and deletions at the ends of a transcript produces exon swap 'ab' and 'ba' type patterns. These events occur at a frequency much greater than expected by chance (Table 2.5). The mechanism by which swaps are generated at 5' ends is different from that at 3' ends. At the 5' end of transcripts, selection of alternative exons must be

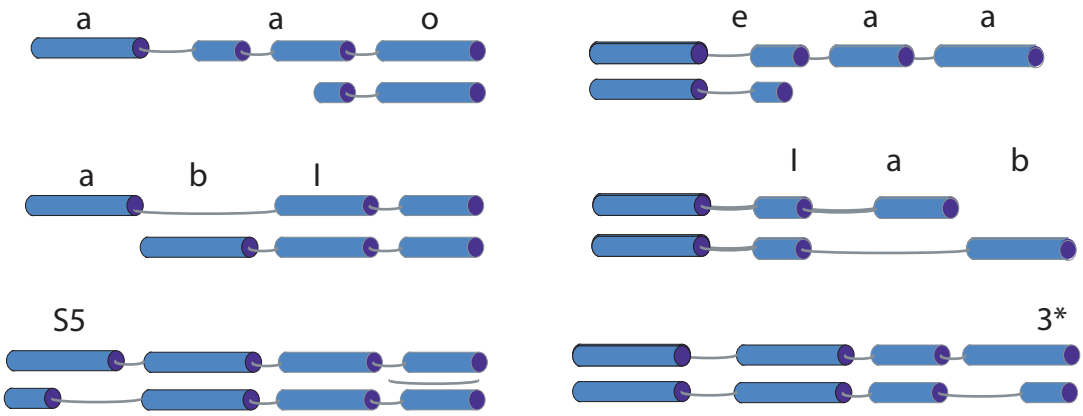

Figure 2.9: Examples of transcription initiation and termination patterns. Patterns of type 'aa' and 'bb' are frequently generated by change in the location of transcription initiation or termination. Highly enriched exon swap patterns 'ab' and 'ba' are generated through coordination of spliceosomes and transcription initiation/termination machinery. Enrichment of the 'S5' pattern indicates linkage between transcription initiation and location of the 5' splice site of the first intron. Enrichment of the '3*' pattern indicates that there is linkage between transcription termination and location of the 3' splice site of the last intron.

done by transcription initiation machinery first, and splicing occurs later, and it is not clear how much of a role spliceosomes play in the process. On the other hand, at the 3' end, selection of alternative exons requires coordination of spliceosomes and transcription termination machinery.

Swap	Observed	Expected	Enrichment
3' swap abI	70	10+/-3	7
3' swap baI	98	9+/-2	10
5' swap Iab	436	45+/-6	9
5' swap Iba	280	47+/-6	6

Table 2.5: Frequency of swaps at the 5' and 3' ends of the transcripts. For simplicity swaps with only 2 exons are show. Swaps at a 5' end are more common that swaps at a 3' end.

The most direct observation of coordination between transcription initiation machinery and the first spliceosomal complex can be seen in the 'S5' pattern, which arises when the first intron has an alternative 5' splice site (Figure 2.9). With an enrichment factor of 5.6 and 120 observations, it is the fifth highest ranked pattern of size two. The implication of this observation is that there exists a mechanism for coordinating location of the 5' splice site in the first intron with the location of the transcription initiation site. Similarly, coordination between the last spliceosomal complex and transcription termination can also be seen in the $\sim$2 fold enrichment of the '3*' pattern.

2.4 Discussion

There are three major mechanisms that can increase transcript diversity: alternative transcription initiation, alternative splicing, and alternative transcription termination. In theory, alternative splicing can generate an enormous number of exon combinations, while alternative transcription initiation and termination events can only modify exons on the edges of a gene. Our study is in agreement with other large scale analyses of cDNA

data [40, 58, 43], showing that the role of alternative transcription events in generating diversity has been largely under-appreciated. We find that exons modified by alternative splicing (A,B,5,3,and X) constitute less than 4% (4143 out of 106,858) of all exons, while exons that are changed through alternative transcription initiation/termination, some times in combination with splicing, (a,b,e,o,y,z), make up 20% (21,573 out of 106,858) of all exons.

The majority of internal splicing patterns are formed by a single spliceosome, choosing a single different splice site. This is the case for the formation of multiple exon skip 'AA' patterns, '5A' patterns, and 'A3' patterns. The only pattern that does not seem to fit this model is the '53' pattern. This pattern is produced when both 5' and 3' splice sites of an intron are changed by a single spliceosome, suggesting that the change in location of splice sites is coordinated. This observation is contrary to the established model of intron removal, where 5' and 3' splice sites are recognized independently by U1 and U2 snRNPs, and these molecules do not form a complex until later ATP driven steps in the spliceosomal assembly pathway [6]. So why do we see enrichment in the '53' pattern? One possible explanation, outlined in Figure 2.8, is that '53' patterns are formed by two splicing reactions of overlapping introns. However, this could only explain approximately 60% of the contributions. A second possibility, is that these are U12 splice sites. However, analysis of motif pattern around '53' introns rules out this possibility. Thus, the mechanism of formation of '53' pattern remains a mystery.

Apart from a small enrichment in exon swaps, there seems to be no evidence for interaction between multiple spliceosomal assemblies. In particular, we do not find evidence for strong linkage between adjacent spliceosomes across an exon, as suggested by the exon definition model of splicing [57]. If adjacent spliceosome complexes communicate across exons, why do we observe only three examples of exons where both ends (pattern 'X') have been modified? Furthermore, other patterns that are formed by adja-

cent spliceosomal assemblies such as '55' and '33' are also infrequent and have the same probability as one would expect by chance (Figure 2.6).

Several enriched patterns imply interaction between splicing and transcription machinery. Interaction between transcription initiation and the first spliceosomal complex frequently result in changes in the 5' end of the first intron ('S5' pattern). Interactions between the transcription termination complex and the last intron's spliceosomal assembly frequently result in a changed 3' splice site for the last intron ('3*' pattern). Exon swaps on both ends of transcripts occur at frequencies much higher than expected by chance ('ab' and 'ba' patterns).

In summary, pattern analysis shows that most changes to the exon structure of transcripts are products of alternative transcription initiation and termination, that the majority of alternative splicing patterns are produced by a single spliceosome working independently of other spliceosomes, and that interaction between spliceosomes and transcription machinery is responsible for generation of many highly enriched splicing patterns.

2.5 Methods

2.5.1 Sources, Quality Control, and Selection of Major Isoform

In this study we limited our analysis to full length cDNA sequences collected from the Unigene, Refseq, and H-inv databases. cDNA sequences were aligned to contigs in the Human Genome, and checked for errors. Major isoforms were selected on the basis of most commonly observed isoforms in EST libraries. See Common Method Appendix for a full description of data sources, quality control checks, and ranking criteria used to select major isoform.

2.5.2 Selection of Minor Isoforms

The exon structure of the major isoform and potential minor isoform were aligned using genomic coordinates, and alignments were filtered using the follow criteria: 1. To eliminate partial cDNA sequences, we removed all transcripts that are missing more than 30% of exons when compared to the major isoform. 2. We removed all transcripts that have identical exon structure with the major isoform. 3. We removed all transcripts that do not have any overlapping exon with the major isoform, and all single exon transcripts. 4. We removed all transcripts that had intron retention events relative to the major isoform.

2.5.3 Random Pattern Generator

The overall process is shown in Figure 2.5. The random pattern generation algorithm consists of three major steps; partitioning of symbols into three groups (5' end symbols, internal symbols, and 3' end symbols), shuffling of symbols within each group, merging of groups, followed by validation of the resulting pattern (see methods). We generated 10,445 random patterns, one for each of the real patterns, and calculated the frequency of all pairs of symbols. The process was repeated 100 times, to obtain the mean and standard deviation for each substring count.

1. For each real pattern we generate a random pattern of the same length. Symbols for the random patterns are drawn from the three buckets. The first bucket consists of symbols describing transcript differences arising from process involving transcription initiation (a,b,o,y), the internal bucket consists of symbols describing differences arising from processes involving only alternative splicing (I,A,B,5,3,X), and the last bucket contains symbols describing differences involving transcription termination (a,b,e,z). The number of symbols selected from each bucket is the same

as in the corresponding pattern for that group. For example, if the real pattern reads 'SaaoIIIIbbb*', we would draw 3 random symbols from first bucket, 4 random symbols from internal bucket, and 3 random symbols from last bucket.

2. Symbols from the three groups are merged into a single pattern.

3. The new pattern is checked to see if it is valid, using the following rules. a. There are can only be one transcription start symbol ('o' or 'y') and only one transcription end symbol ('e' or 'z') per pattern.

   b. Mixtures of 'a' and 'b' before symbols 'o' or 'y' are not allowed.

   c. Mixtures of 'a' and 'b' after symbols 'e' or 'z' are not allowed.

   d. Combinations 'aA', 'aB', 'bA', 'bB', 'SA', 'SB', 'A*', 'B*' 'S3' and '5*' are not allowed.

4. If a pattern fails, reshuffle within groups and validate again.

   a. If a pattern fails more than 100 times, try swapping 'a' and 'b' symbols between the start and end groups.

   b. If a pattern fails more than 200 times, abandon the pattern and go back to step 1.

Chapter 3

Chapter 3. Noisy Splicing

3.1 Abstract

For some time, the number of known alternative Human isoforms has been increasing steadily with the amount of available transcription data. To date, over 100,000 isoforms have been detected in EST libraries, and nearly 90% of Human genes have at least one alternative isoform. In this chapter, we propose that most alternative splicing events are the result of noise in the splicing process. We show that the number of isoforms and their abundance can be predicted by a simple error model that takes into account two factors: the number of introns in a gene and the expression level of a gene. Furthermore, we show that there is substantial selection pressure to reduce the frequency of alternative splicing in highly expressed genes and genes with many introns. We argue that these observations are consistent with error rates tuned to reduce the toxic effect of accumulation of misfolding proteins in the cell and to ensure that an adequate level of functional product is produced. Based on simulation of sampling of virtual cDNA libraries, we estimate that the average Human cell, with 800,000 transcripts, may contain as many as 100,000 alternative isoforms due to splicing noise.

3.2 Introduction

The number of Human genes with alternative splicing is presently not well established. Early estimates based on EST data suggested that around 35-40% of all genes

have at least one alternative isoform [18, 59]). Current estimates based on a larger collection of EST libraries and microarray experiments show numbers as high as 90% (Figure 3.1, [21]). It is now clear that nearly every gene with potential for splicing produces alternative isoforms. These observations raise an important question: are these isoforms functional or are they in some sense the product of stochastic noise in the operation of the splicing machinery?

In response to specific environmental and cellular conditions, transcription and splicing machinery might generate a wide diversity of unusual transcripts, and in any particular case it is difficult, if not impossible, to determine the importance of each transcript without appropriate experiments. Nevertheless, the question of the overall impact of alternative splicing on function can be addressed using statistical methods.

Numerous bioinformatics studies have analyzed tissue specificity, species conservation, domain architecture, sequence properties, and structural properties of isoforms [59, 60, 39, 61, 10]. Most studies relate the probability of an alternative splice isoform having function to tissue specificity, abundance, or conservation across species. It is estimated that approximately 10-20% of all of alternative splicing events are conserved across two or more species [27, 25, 62, 64, 63]. Conserved alternative splicing events are found to be enriched in characteristics consistent with generation of novel molecular function, such as increased coding frame preservation, increase in abundance, and preference for changes in the functional regions. While some of these likely have function, it is by no means clear that all do. Additionally, the functional properties of the much larger set of low abundance species specific isoforms is left open.

There are essentially three hypotheses that can explain the presence of these isoforms. 1. Alternative isoforms produce novel protein sequence and thus generate new functionality [22, 23, 65, 60] 2. Alternative isoforms that do not code for functional proteins may regulate total abundance of functional isoform(s) by nonsense mediate decay

(NMD) or protein degradation pathways [21, 66]. 3. Alternative isoforms are the result of stochastic noise in the splicing process [23]. They are unlikely to code for functional protein products, but as long as they do not negatively impact the normal function of a gene there is little selection pressure to limit their production. It has been proposed that alternative isoforms might serve as a testing ground for molecular evolution [67, 32, 68].

Obviously, potential long-term evolutionary benefits of alternative splicing are not the cause of the observed diversity of isoforms. In the short term, the critical aspect of the system that determines isoform diversity is the error rate of splicing machinery. In turn, the error rates are determined by selection pressure from such factors as the energetic and toxic consequences of errors, as well as the requirement of producing the minimum quantity of active transcripts required for biological activity.

In this chapter, we explore the consequences of the idea that errors in splicing largely determine the number of alternative isoforms and their transcript abundance. The key observations supporting the noisy splicing hypothesis are that the number of isoforms increases as a function of the expression level of a gene and the number of introns in a gene. The greater the number of splicing reactions - the greater the number of opportunities for a mistake - the greater the number of isoforms produced. We find that there is large variability in error rates and that genes with many splicing reactions have reduced error rates. Based on these observations we propose that there is selection pressure on highly expressed genes and genes with a large number of introns to maintain low levels of alternative splicing.

As a test of the hypothesis, we have developed three models of error rate per splicing reaction: 1. A constant error rate; 2. Error rates varying with the number of introns in a gene; 3. Error rates varying with the number of introns and transcripts of a gene. Each model was tested by simulating the production and experimental sampling of transcripts from virtual cDNA libraries. The observed data are most consistent with the error model

that takes into account the number of introns and the relative abundance of a gene. Furthermore, we find that the density of predicted Exon Splicing Enhancers increases with the number of splicing reaction, implying better determined splice sites in genes undergoing many splicing reactions. The success of the model in reproducing observed trends in the experimental data strongly supports the view that a large fraction of minor isoforms is indeed non-functional.

3.3 Results

3.3.1 Overview

Before describing the results, it is useful to clarify some basic definitions used in this study. First, we define the major isoform of a gene as the isoform that is most commonly observed in EST libraries. Using the major isoform as a reference, we define an alternative intron as an intron that differs at the 5' and/or 3' splice site from corresponding intron in the major isoform. If a transcript of a gene contains one or more alternative introns, we call it an alternative transcript. An alternative isoform is defined as a unique splicing pattern that is different from the splicing pattern in major isoform. By definition, a single such isoform can be represented by multiple transcripts.

Figure 3.1 shows the distribution of number of alternative isoforms per gene derived from the Complete Set of 8674 EST libraries (see Datasets). Nearly 90% of all genes have alternative splicing, and the majority of genes have three to six alternative isoforms. Of course, given that present EST libraries sample only a small fraction of transcript space, only a fraction of all isoforms have so far been observed. Some of these detected isoforms have been produced through regulated selection of splice sites, while others may be the product of stochastic noise [23]. More precisely, assuming that splicing machinery

does not select splice sites with perfect accuracy, one would expect that some fraction of alternative isoforms have been produced in error.

If mistakes occur, then they should happen at low frequency and in turn this implies that the fractional abundance of alternatives should be low. To get an approximation to the fractional abundance of alternative transcripts, for each gene we calculate the fraction of all EST sequences that had alternative introns. The histogram of fractional abundance is shown in Figure 3.2. Indeed we find that the large majority (80%) of all alternative introns are present at less than 12% fractional abundance (mean 8% , median 4%).

The observation of a large number low abundance isoforms suggests that errors do contribute very significantly to isoform production. A key question is then what is the relationship between error rates and observations? A basic expectation of any error model is that the number of mistakes is a function of the total number of opportunities to make mistakes. For spliceosomes, the number of opportunities is equal to the number of splicing reactions - the total number of introns removed from all transcripts per unit time. Thus, one would expect that number of observed unique isoforms would increase with increasing expression level and with the number of introns in a gene.

The increase in the number of unique isoforms as a function of the number of introns and the number of sampled ESTs is shown in Figure 3.3. Consistent with the noise hypothesis, it can be seen that both quantities contribute to an increase in the number of isoforms. However, it is possible that this increase is not due to noise. First, it could be that greater functional diversity is produced by genes with more introns, and that is what is being observed here. Second, greater EST sampling should result in increase discovery of alternative isoforms, regardless of the mechanism by which these isoforms were generated, be it noise or regulated selection of splice sites.

Furthermore, although one would expect that the number of ESTs per gene is re-

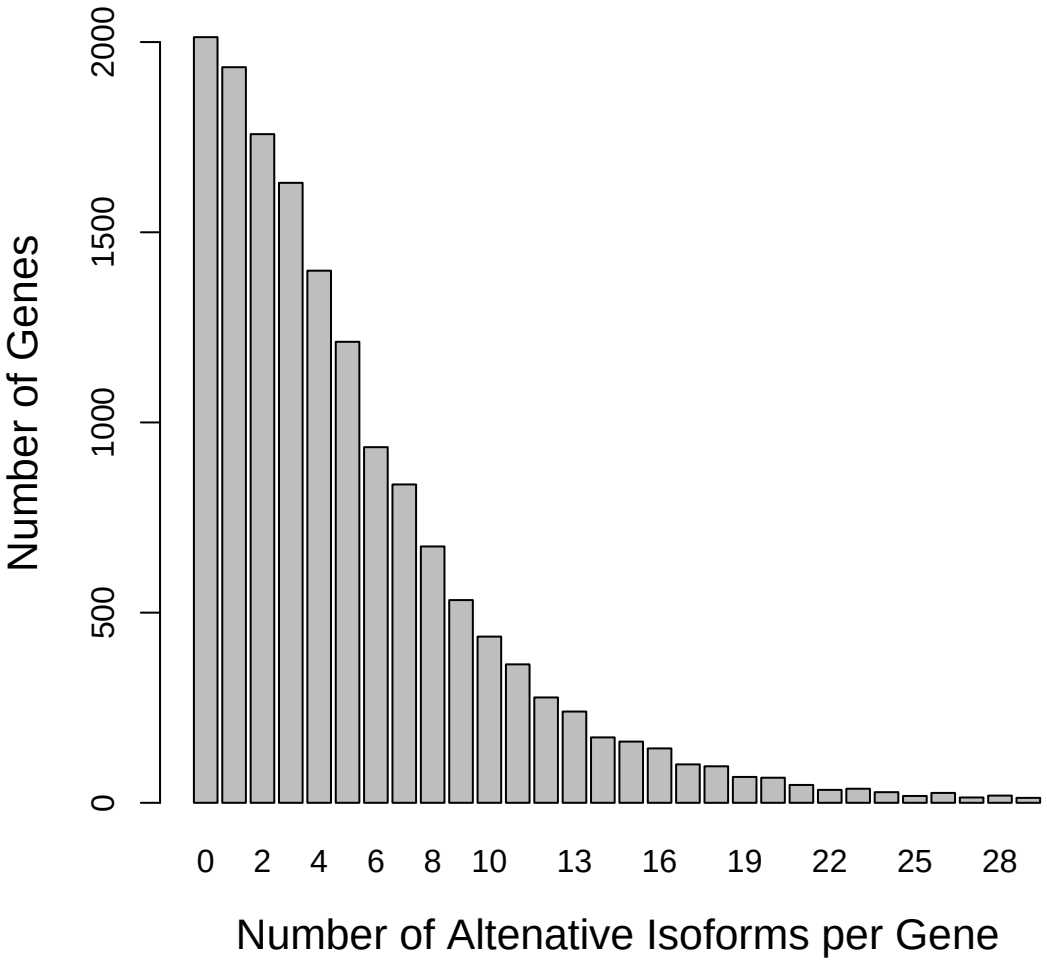

Figure 3.1: Isoform distribution. Distribution of number of alternative isoforms per gene derived from all 8674 Human Unigene EST libraries (15,342 genes $\sim$ 5,313,000 EST sequences). The first bar contains the 2013 genes (13%) with no alternative isoforms. The median number of isoforms per gene is 4.

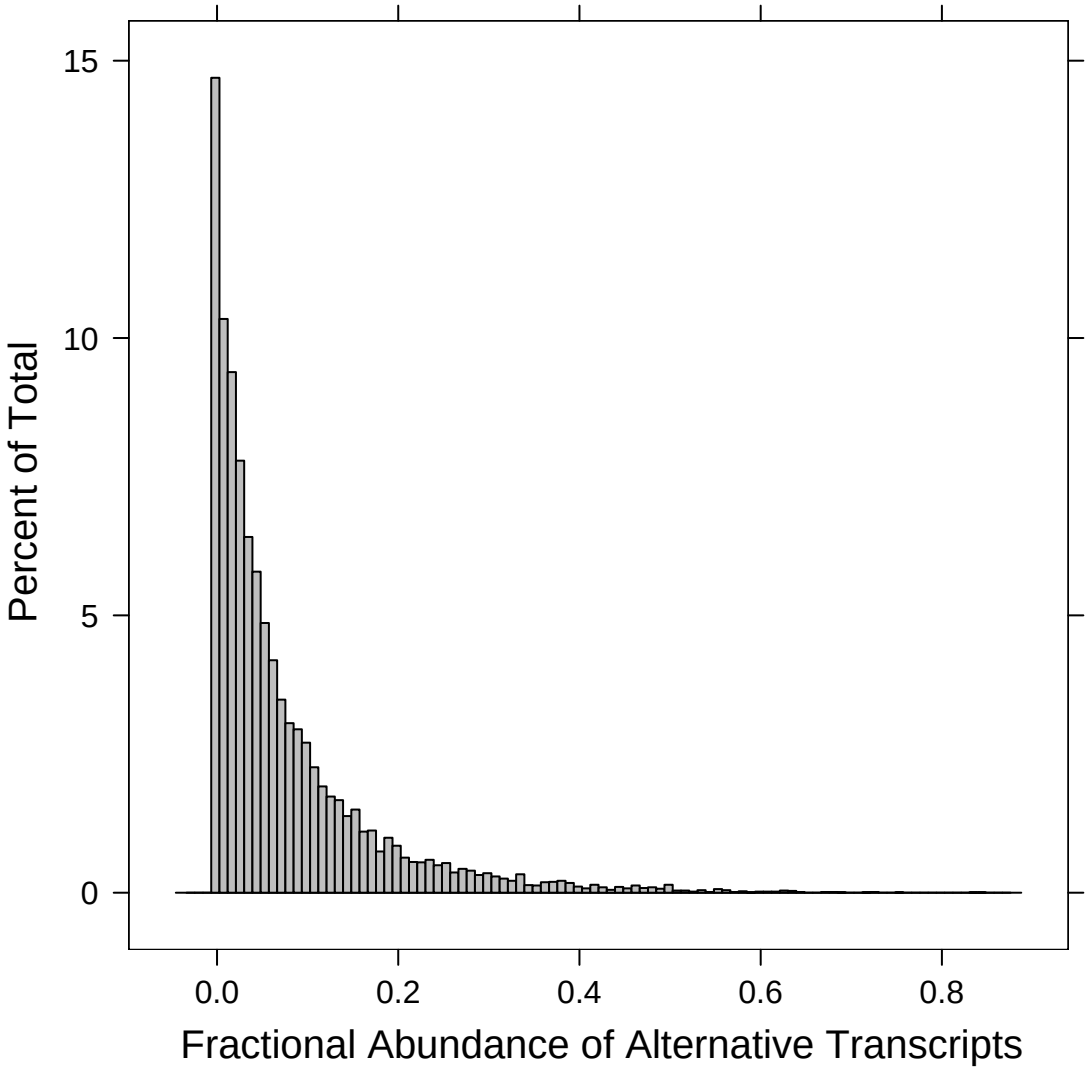

Figure 3.2: Fractional abundance of alternative transcripts. For each gene in the Complete set (15,342 genes), EST sequences of a gene were compared to the major isoform to identify alternative splicing events (see methods). We then calculate the fractional abundance of alternative transcripts as the total number of ESTs with alternative introns divided by the total number of ESTs. The large majority (80%) of alternatives are observed at less than 12% fractional abundance. Median fractional abundance of alternative transcripts is $\approx 4.1\%$.

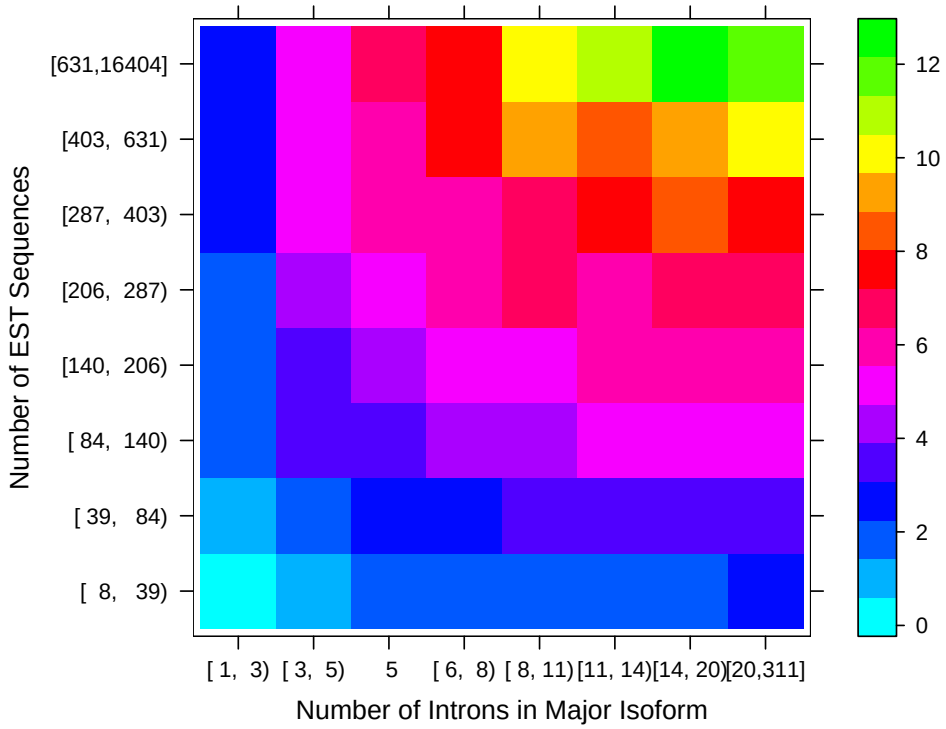

Figure 3.3: Increase in number of isoforms as a function of number of introns and EST observations. Genes from the Complete Set were divided into a 10 by 10 matrix according to number of introns in the major isoform and the number of observed ESTs per gene (each group contains ≈ 150 genes). The mean number of isoforms was calculated for each matrix element. As can be seen in the plot, the number of isoforms increases as a function of both the number of introns per gene and the number of sampled ESTs per gene.

flective of the actual number of transcripts per gene in the cDNA pool, this need not be the case. EST libraries are frequently enriched for rare transcripts through normalization and subtraction procedures, and so the number of observed transcripts are not reflective of actual abundances [69]. There are also possible problems with EST libraries constructed from pathogenic tissues, which might contain many abnormal splicing events. Before error rates can be estimated, these issues need to be resolved.

The problem of limited sampling of ESTs can be addressed by the use of simulations, as described later. The problem of normalized, subtracted, and pathogenic tissue libraries can easily be addressed by removal of all such libraries from the analysis. Thus, in addition to Complete Set of all 8674 EST UNIGENE libraries [70], we created three EST library subsets: the CGAP Subset of 325 non-normalized libraries derived from normal tissue samples [71], the CGAP Lung Subset of 16 libraries derived from a normal lung tissue, and the single UNIGENE EST library derived from normal pancreatic islet cells (HS Lib8840 [72]). In all four sets, we find that trends predicted by noisy splicing.

3.3.2 Estimation of Error Rates from Observed Data.

It is possible to calculate an implied splicing error rate per splicing reaction for a set of genes directly from observed data, using the assumption that most alternative introns are the result of mistakes in selection of splice sites. If errors occur at a constant frequency then the number of alternative introns produced should grow linearly with increase in the total number of removed introns. Figure 3.4A shows the average number of detected alternative splicing reactions (that is, the number of observed alternative introns) as a function of the total of observed splicing reactions (the number of detected introns in all EST sequences of a gene). As expected, the number of detected alternative reactions increases with increasing reactions, but, surprisingly, the increase is nonlinear. Figure

3.4B shows implied error rate per reaction, defined as the average ratio of number of detected alternatives reactions over the total number of splicing reactions, also as a function of the number of splicing reactions. We find that genes that undergo more splicing reactions make relatively fewer mistakes, that is they have lower error rates. This is by far the most surprising observation in our analysis. The decline is not due to sampling biases, since the number of detected alternative introns and the total number of detected introns are subject to exactly the same biases.

Based on these observations we propose that selection pressures influence error rates in two primary ways. Genes with many introns must reduce error rates if adequate quantities of normal protein products are produced, since it is otherwise unlikely that all introns could be removed without producing at least one mistake. For example, with a 2% error rate, nearly all transcripts of a gene with 100 introns will contain at least one error ($0.98^{100} \approx 13\%$). Conceptually, it is also easy to see why genes with large abundance must reduce error or risk toxic effects on the cell. Production of large quantities of misfolded protein products may overwhelm chaperones, and cause toxic protein aggregation [73, 74].

Thus, we propose that there are two key components that influence error rate: the number of introns per gene and the expression level of a gene. However, the impact of each component on error rates cannot be determined directly from EST data. First, only a fraction of all introns is sequenced by ESTs. Second, only a small fraction of all transcripts is sampled by ESTs. In the next sections, we address these issues by simulations that take into account these biases.

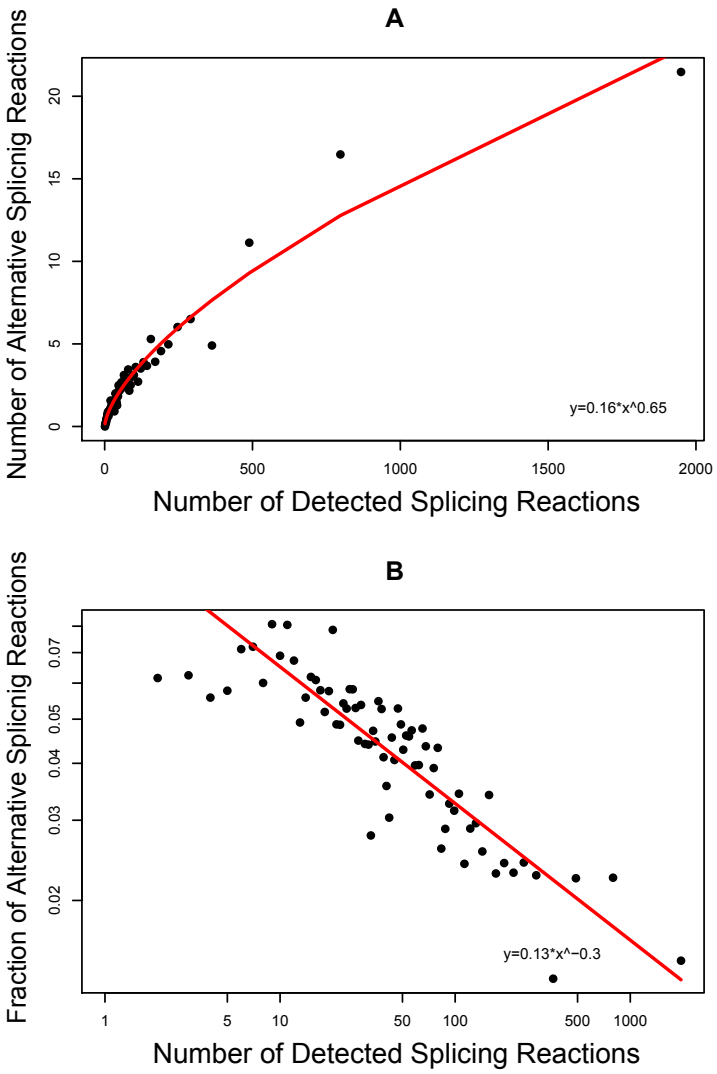

Figure 3.4: Estimates of error rates per splicing reaction from EST sequences. The number of detected splicing reactions is a count of number of all introns that have been observed in all EST sequences of a gene. The number of alternative splicing reaction is a count of introns that differ in 5' and/or 3' splice site from corresponding intron in the major isoform. Genes in CGAP subset were divided into ≈ 100 equal size groups based on number of splicing reactions. Within each group, the mean number of splicing reactions and the mean number of alternative splicing reactions were calculated. As can be seen, the increase in number of alternative introns is nonlinear. Panel B: approximate error rate estimated as fraction of all splicing reactions that are alternative. Genes with many splicing reactions make fewer mistakes producing a decreasing fraction of alternative introns.

3.3.3 Overview of Error rate and Sampling Simulations

We developed three models of error rate per splicing reaction. The first model assumes that the error rate is the same for all genes. The second model assumes that the error rate is a function of the number of introns in a gene. The third assumes that the error rate is a function of the number of transcripts and the number of introns for a given gene. The error models are used as input to a virtual transcript simulator, which generates the transcript contents of a cDNA library, consistent with the error assumptions. We then simulate experimental EST sampling from this cDNA library, to create a virtual EST library, which can be compared directly to real EST libraries.

Experimental cDNA libraries contain transcripts from millions of cells, and each cell contains approximately 800,000 transcripts. No two cells are identical in their transcript content and most (40-48%) transcripts are present at abundance levels of less than 1 copy per cell [2]. To generate a virtual cDNA library we require three inputs: the number of introns in each gene, the absolute message abundance (transcripts per cell) for each gene, and a detailed error model. We assume that the major isoform of a gene is produced most frequently, and take the intron count directly from the Refseq full-length cDNA. To get an approximate number of transcripts per cell, we convert microarray values into copies per cell, after calibration with ESTs per gene (see Methods). Based on approximate copies per cell, intron count, and the choice of one of the three error models, we simulate the transcript content for 10000 cells using the virtual transcript simulator. Although memory limitations do not allow us to simulate a larger number of cells, we can show that increasing the number of cells does not significantly affect the outcome.

Each virtual transcript is represented as an intron binary pattern, where '0' indicates that both boundaries of an intron are as in the major isoform, and '1' represents an alternative splicing event where one or both boundaries are different. For each gen-

erated transcript, at each exon/intron junction, the simulator either maintains the major isoform boundary (a '0'), or a splicing error causing a boundary change is introduced (a '1'), with a probability determined by the characteristics of the particular model. Once all transcripts in the set of cells have been generated, we mimic the cloning step and then the sequencing steps in the EST experiments. For this purpose, we randomly pick the same number of virtual ESTs from the generated cDNA library for a gene as were observed in real EST experiments, and truncate them to include the same number of introns as observed in real EST sequences (See Figure 3.11 and Methods for further details).

We used the CGAP Library subset and the Lib8440 library as sources of real EST data. Our findings for the CGAP Library subset are summarized below. The findings for Lib8440 are in qualitative agreement with the CGAP sample and are included as supplementary data.

3.3.4 Model 1: Constant Error Rate

The simplest model of noise assumes that splicing machinery makes mistakes at a constant error rate 'p' per splicing reaction. In this model, all introns are equivalent - that is, the error rate is the same for all introns regardless of gene, number of introns, transcript abundance, intron length, splice site strength, or any other factors. Ten values of p were tested starting at 1% and ending at 10%. As expected from Figure 3.4, none of the p values produced good fit to observed data. The result with p=3% per splicing reaction is shown in Figure 3.5.

As dictated by the fixed error rate, the model produces an approximately constant fraction of alternative splicing reactions as a function of total number of splicing reactions (Panel A), whereas the observed data falls steadily. Not surprisingly, the model predicts a raise in number of alternative isoforms with increase in number of splicing reactions

(Panel C). The model correctly predicts distribution of the number of alternative isoforms per gene (Panel B). However, none of the simulations accurately reproduces the slopes of the experimental curves. The simulation shows an increase in the fractional abundance of alternative transcripts with an increase in the number of splicing reactions (Panel D), while the observed data is approximately flat. It is quite evident that this model is a poor fit to observed data.

3.3.5 Model 2: Error rate Dependent on the Number of Introns

Consider two genes, both transcribed at the same level of 100 transcripts, where the first gene has 10 introns and the second gene has 100 introns. Assuming a 1% splicing error rate for both genes, the first gene would nearly always (90% of the time) produce the major isoform, while the second gene would produce the major isoform only 37% of the time.

In Model 2, we assume that selection will act to reduce this large effect. In this model, genes with many introns will have a lower error rate per splicing event compared to genes with few introns, with the error rate tuned such that on average, a fixed fraction α of all transcripts of each gene are alternative. Given α , the error rate per splicing reactions 'p' for a gene with N introns is obtained from Equation 1.

$$p = 1 - (1 - \alpha)^{\frac{1}{N}} \tag{3.1}$$

Figure 3.6 shows the result of simulations with a best-fit parameter α=0.2. Inclusion of intron counts in the error rate calculation results in a small improvement compared to constant error rates. As can be seen in Panel A, at a low number of splicing reactions, there is an initial decrease in error rate as a function of number of splicing reactions,

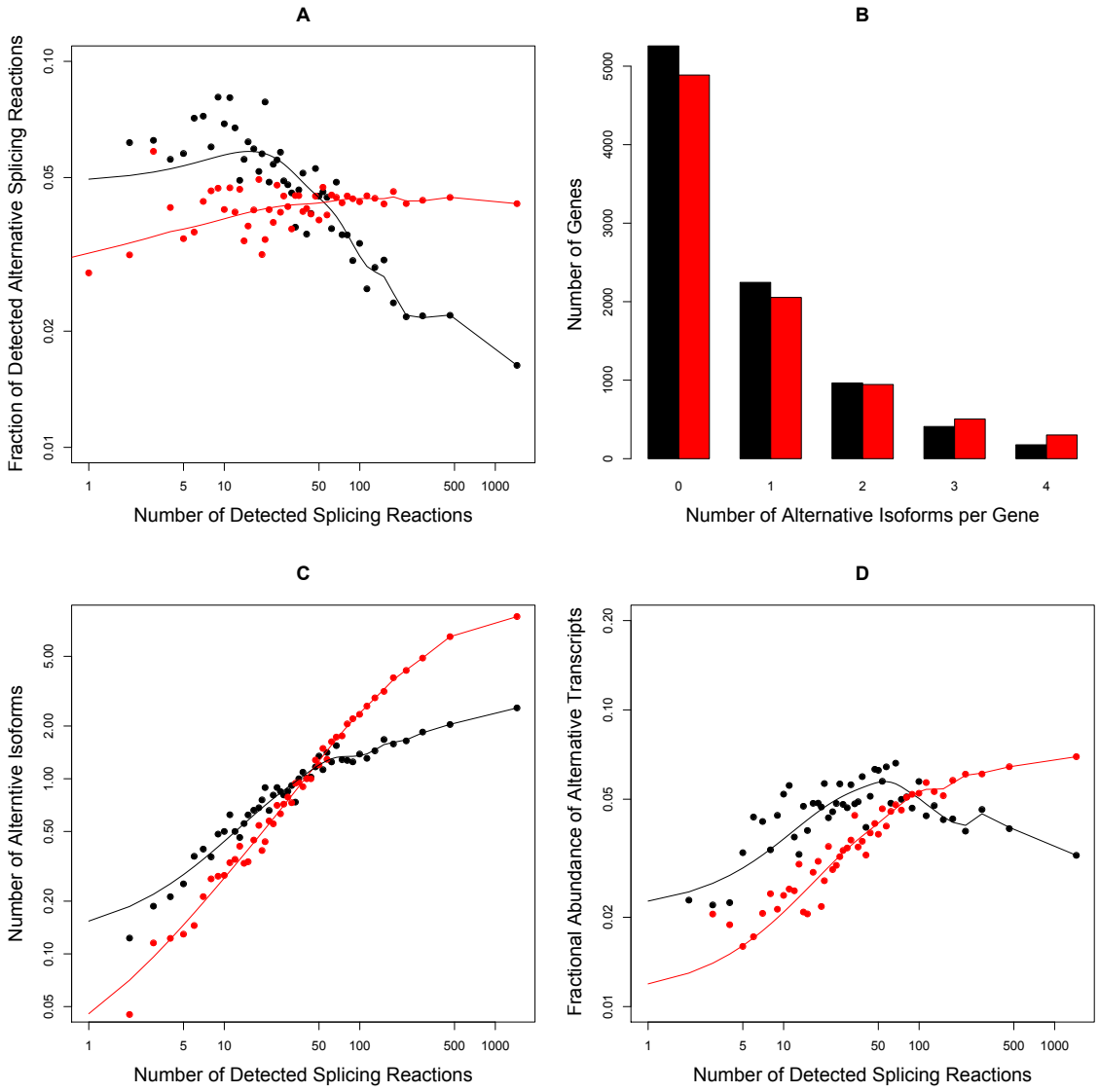

Figure 3.5: Model 1. Simulation of sampling in a virtual cDNA library with 10000 cells. Transcripts generated with a constant error rate. Red points - simulation result with error rate of 3%. Black points - observed data in the CGAP Library Subset. A: Error rate simulated by model compared to observed implied error rate. B: Number of Detected isoforms per gene distribution. C: Increase in number of detected isoforms as a function of number of detected splicing reactions. D: Fractional abundance of alternative transcripts. With the exception of number of isoforms per gene (panel B), this model is a poor fit to observed data.

consistent with the observed data. However, at high values (>100) the simulated error rate rises, while the observed values continue to decline. This model is also a better fit to the number of detected isoforms at a low number of splicing reactions (Panel C) and the fractional abundance of alternative transcripts (Panel D), but fails thereafter.

3.3.6 Model 3: Error Rate Determined by the Number of Introns and Transcript Abundance.

In Model 3, we propose that the error rate per splice junction is a function of both the number of introns and the number of transcripts. The assumption here is that selection pressure tends to limit the total number of noise transcripts produced by all genes, since these will produce non-folding protein products that will saturate the chaperone machinery and/or aggregate, and so be toxic [75, 74].

The error rate per splicing reaction function assumes the same form as in equation 1, with the additional contribution from the total number of non-major isoforms produced:

$$p = 1 - (1 - \frac{\alpha}{1 + \beta * T})^{\frac{1}{N}} \tag{3.2}$$

where T is the number of transcripts per cell, N is number of introns, and α and β are two constants which determine the influence of the number of introns and the number of transcripts on the error level. When $\beta = 0$, the model becomes equivalent to model 2, where the error rate varies only with the number of introns. The higher the value of β, the more influence from the toxic effect of many noise transcripts. We used a grid search of α between 0 and 0.5 and β between 0 and 0.3 to find the combination of parameters which produced the best fit to the observed data.

We find that α ranging from 0.35 to 0.45 and β ranging from 0.01 to 0.02 produces

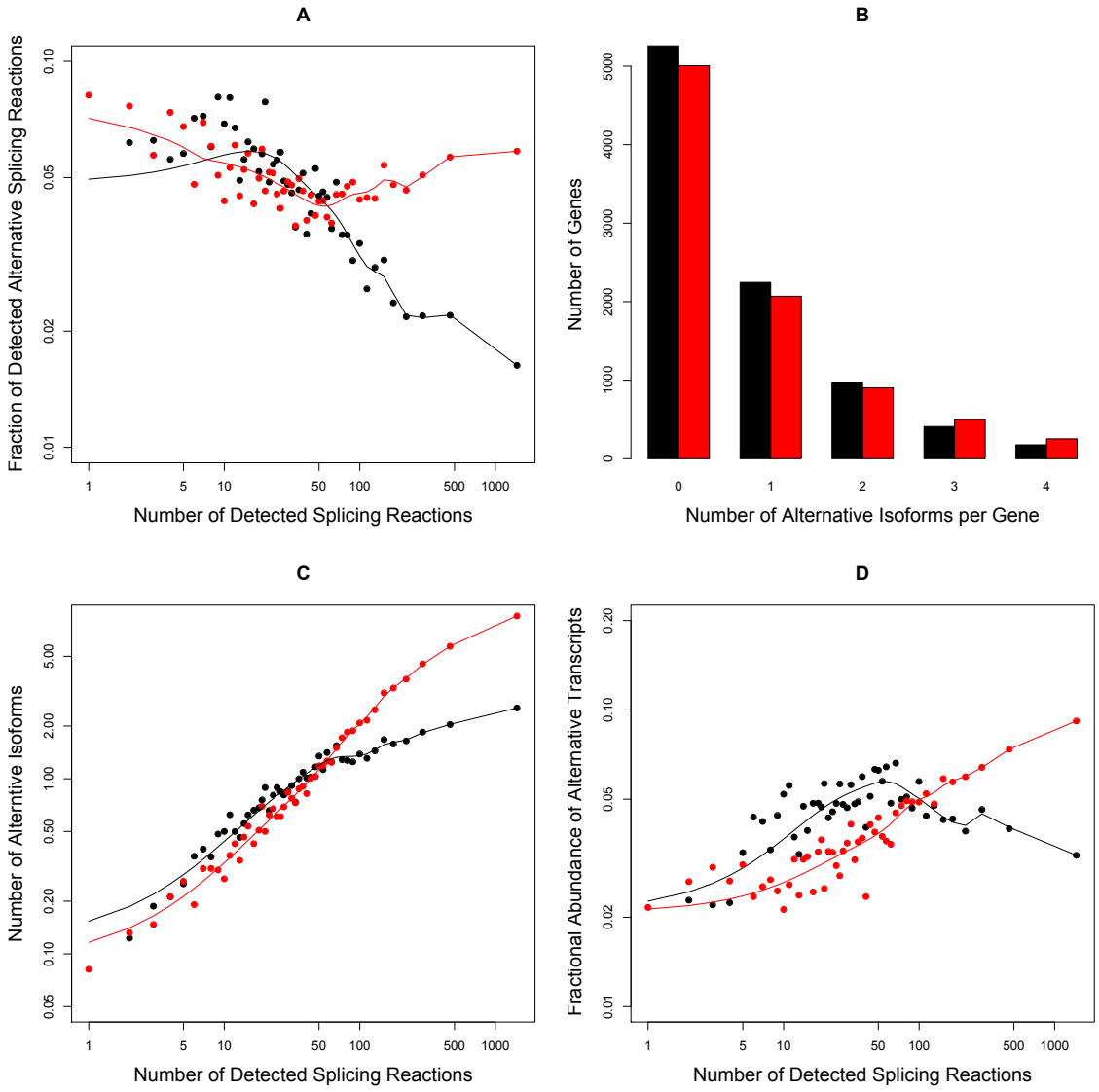

Figure 3.6: Model 2. Simulation of sampling in a virtual cDNA library with 10000 cells. Transcripts are generated with an error rate implied by Equation 1 and $\alpha = 0.2$. Red points- predicted data. Black points - observed data in the CGAP Library Subset. A: Error rate simulated by model compared to observed implied error rate. B: Number of Detected isoforms per gene distribution. C: Increase in number of detected isoforms as a function of number of detected splicing reactions. D: Fractional abundance of alternative transcripts. Model 2 produces better fit to observed data at low number of splicing reactions, but fails for high ($>$100) numbers of splicing reactions.

a good fit. Figure 3.7 shows the results of simulations with $\alpha = 0.4$ and $\beta = 0.015$. Figure 3.7A shows that inclusion of abundance corrects the problem with model 2, reproducing the observed decline in estimated error rate throughout the entire range of splicing reactions. Figure 3.7D shows that model 3 also correctly reproduces the nearly constant fractional abundance of alternative transcripts, although the predicted fraction of alternatives is a few percentage points lower than observed in real EST libraries. We also observe that model 3 over-predicts number of isoforms for genes with many (> 100) splicing reactions. We could further refine the parameters but it is not clear if this is useful, given the other approximations in the model. The goodness of fit between observed data and all three models is shown in Table 3.1.

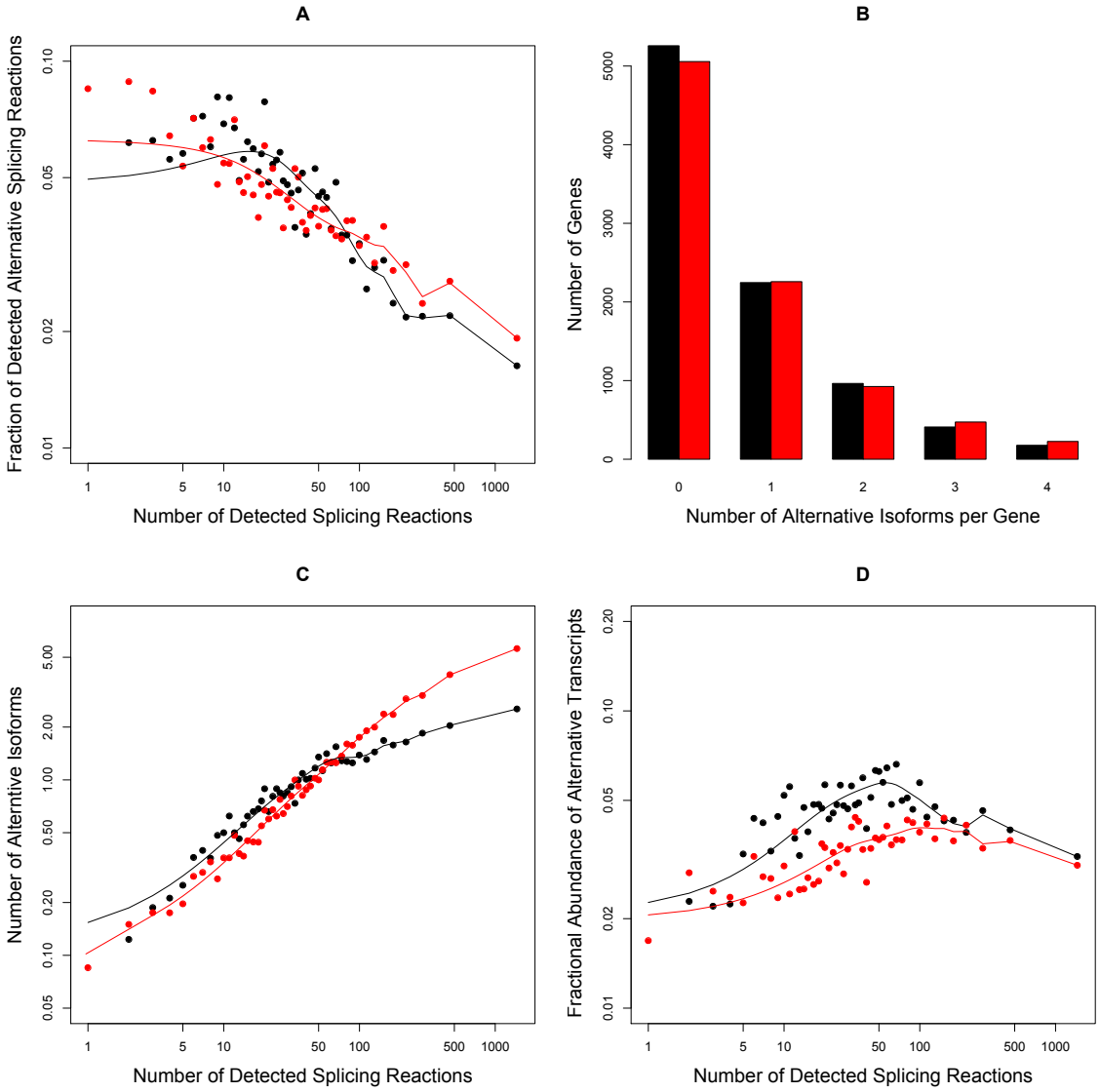

Figure 3.7: Model 3. Simulation of sampling in a virtual cDNA library with 10000 cells. Transcripts generated with error rate implied by Equation 2 with parameter values $\alpha = 0.4$ and $\beta = 0.015$. Simulation in red, observed data in black (CGAP Subset). A: Error rate as simulated by model compared to observed implied error rate. B: Number of Detected isoforms per gene distribution. Panel C: Increase in number of detected isoforms as number of detected splicing reactions. D: Fractional abundance of alternative transcripts. Model 3 correctly reproduces the decrease in error rates, number isoforms per gene, and fractional abundance of alternative transcripts. It over-predicts the increase in number of isoforms for genes with high (>100) numbers of splicing reactions.

3.4 Discussion

It is now clear that nearly every Human gene has multiple isoforms, and more isoforms are discovered with every new EST library. Is it the case that we are discovering new biological activity or could it simply be that we are observing accumulation of occasional mistakes in the selection of splice sites by splicing machinery - splicing noise?

There is no doubt that some fraction of all isoforms are functional. It is estimated that the fraction of all alternative splicing events that are conserved between two or more species is approximately 10-20% [27, 25, 62, 63, 64]. Numerous bioinformatics and microarray based studies have found that isoforms conserved across species tend to preserve coding frames [31, 39], are less frequently subject to nonsense mediated decay [39, 35], and are expressed at higher abundance. It should be noted that our knowledge of conserved splicing events is biased toward genes that are expressed at high abundance, because those are more likely to get sampled by EST transcripts [26]. We do not actually know what fraction of undetected alternative events will be conserved. Nevertheless, assuming that function and conservation are correlated quantities, what is the explanation for the much larger set of species specific isoforms?

The hypothesis advanced in this chapter is that the majority of isoforms are products of noisy splicing. Our argument is based on the observation that the number of alternative isoforms increases as a function of two quantities: total expression of a gene and number of introns in a gene. Simply put, the more frequently introns are removed and the greater the number of isoforms produced, the more chances there are of making mistakes. Of course this argument on its own is not sufficient proof of this hypothesis.

There are other lines of evidence that support the hypothesis. First, as noted previously, only a small fraction of alternative isoforms are found in two or more species and most isoforms (more than 70%) do not show clear tissue specificity [28, 29]. Sec-

ond, a large fraction (34%) are expected to be subject to nonsense mediated decay [33]. Third, although there will be exceptions, one would expect that in most cases the level of expression required to perform an alternative function would be similar to that required for the major isoform. Yet, the level of expression of most alternative isoforms is very low (median 4%) compared to the corresponding major one (as can be seen in Figure 3.2). Fourth, examination of the implied protein sequences and structures of alternative isoforms shows that in most cases the structures are non-viable (Chapter 4).

The idea of splicing noise is not novel, and has been suggested by several researchers [23, 20, 38, 76]. It has been assumed that error rates of splicing machinery are very low, and that if spliceosomes make mistakes, these mistakes would represent only a small fraction of all observed isoforms [23]. For example, Kan et al. estimated error rates to be less than 0.01 per splice junction [38]. However, development of error rate models was not a major focus of that study. More recently, Neverov et al. [76] proposed a constant error rate model with a frequency of 0.012 per splice junction. Similarly to this study, the model was used to simulate isoform production, but not with the explicit purpose of estimating error rates.

Assuming that splicing machinery does make mistakes, what factors are expected to most influence the error rate? Genes with many introns can not tolerate high error levels because this would result in significant loses of major product. The cell cannot tolerate highly expressed genes having a high error rate because the resulting large number of non-folding protein products would be toxic, either by over-whelming the chaperone system or by forming aggregates. Based on these assumptions, we would expect that genes with many introns and high abundance would be tuned to reduce error rates.

There are a number of possible mechanisms by which error rates may be tuned, such as 'stronger' splice site motifs, an increase in the number of exon/intron splicing enhancer motifs, and an increase in the number of exon/intron silencer motifs. As a simple check of

the hypothesis, we divided genes into 10 groups based on the number of splicing reactions, and computed the average splice site consensus score and the average number of predicted exon splicing enhancer (ESE) motifs in each group. We found that genes with many splicing reactions contain more ESEs compared to genes with few splicing reactions. We did not find the same trends in splice site scores - the scores remained nearly constant for all gene groups. The contribution of each factor to the increase in splicing fidelity will clearly require a more detailed investigation. Nevertheless, the correlation of error rates with ESE density provides strong support for the tuned error rates model.

We argue that reduced frequency of alternative splicing in highly expressed genes is consistent with the hypothesis that error rates are tuned such that an adequate amount of functional product is produced and the overall impact on cells is nontoxic. The latter point is analogous to the arguments by Drummond et al. [77] to explain increased selection pressure against mutations in highly expressed genes. They assert that the explanation for this phenomenon is that there has been significant selection against the accumulation of miscoded proteins, because of their potential toxic effects.

To address problems with sampling biases in EST libraries, and test various assumptions about error rates, we ran a number of computer simulations of EST sampling from virtual cDNA libraries. We tested a constant error rate (model 1), an error rate subject to intron count (model 2), and an error rate subject to intron count and transcript abundance (model 3). We show that the model that takes into account both introns and abundance has a best fit to the observed data. The error per splicing reactions as predicted by model 3 over a wide range: between 0.1 to 9% for genes with 5-20 introns and 1-1000 transcripts per cell(Figure 3.8).

Although this model fits the data, that is not a final proof of correctness. There are several limitations in our simulations. First, we treat all alternative splicing events as errors, and that is clearly not the case. Eliminating the approximately 10-20% of conserved

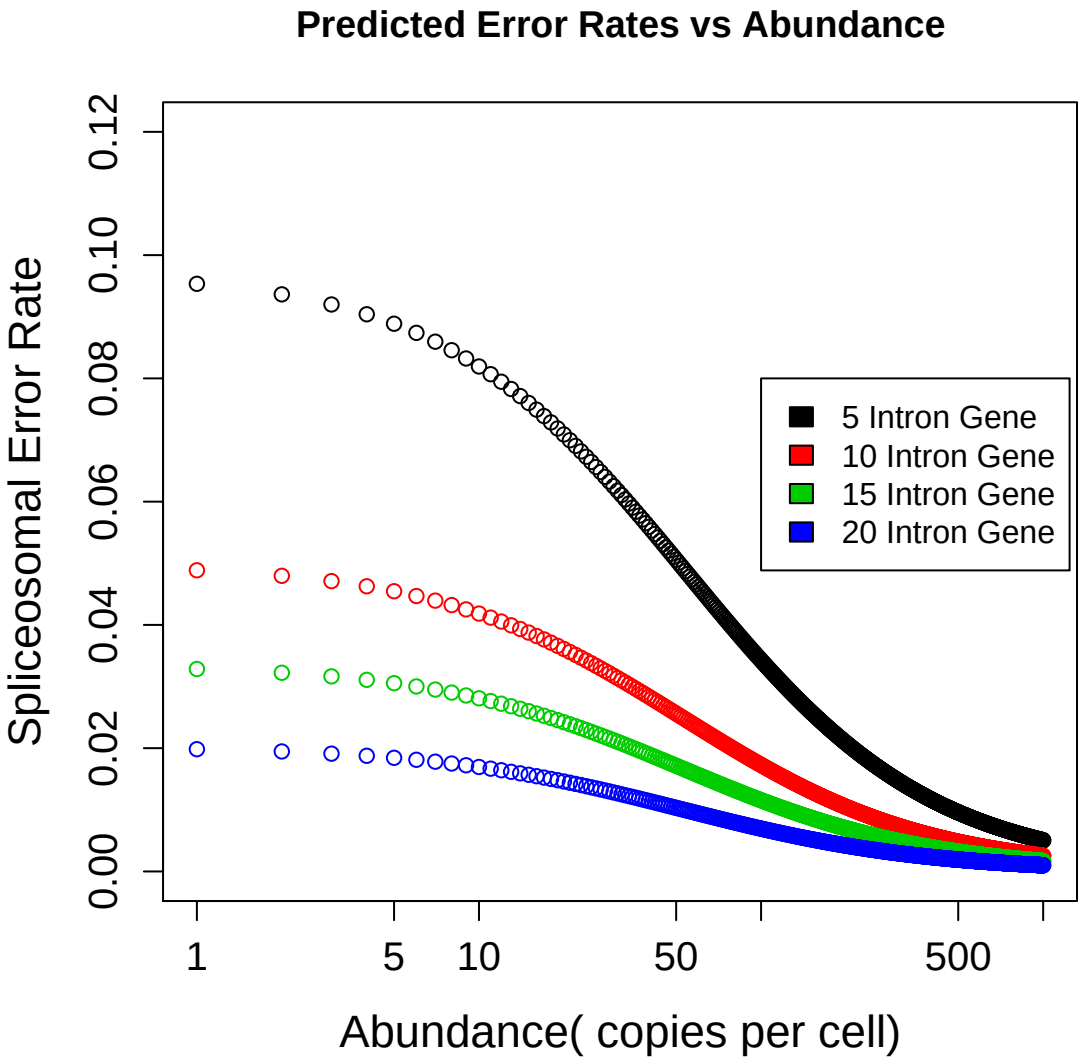

Figure 3.8: Variation in average error rate per splice junction as a function of transcript abundance, for genes with different numbers of introns. Data produced by model 3, with $\alpha = 0.4$ and $\beta = 0.015$. At low abundance levels, genes with few introns are predicted to have high average error rates ($\sim 9\%$), while genes with many introns have low values ($\sim 2\%$), reflecting the greater number of splicing reactions per transcript. At high abundance levels, all error rates are predicted to be low (less than 1%) because of selection against producing a large number of nonfunctional transcripts.

splicing events from datasets would result in slightly smaller error rate estimates. Second, our models do not include information about the strength of splice sites, distances between intron ends, enhancer and silencer motifs, and other factors that influence selection of splice sites. All these factors can be incorporated into simulations, but at cost of increased complexity of interpretation. In spite of the limitations, we are not aware of other possible explanations for the surprising decrease in the fraction of alternative introns with increasing number of introns in a gene and with expression rate.

3.5 Methods

3.5.1 Sources, Quality Control, and Selection of Major Isoform

In this study both DNA and EST sequences collected from the Unigene, Refseq, and H-invDB databases were used. Sequences were aligned to contigs of the Human Genome, and checked for errors. Major isoforms were selected on the basis of the most commonly observed isoform in all EST libraries. See Common Methods Appendix for a full description of this process.

3.5.2 Datasets

Complete Set: EST sequences from all Unigene EST libraries (8674 libraries total) that have a unique mapping to a Refseq gene entry. The Refseq gene must a have mapping to a unique genomic locus. The dataset contains 15,342 genes with 5,313,618 EST sequences that have passed quality control checks.

CGAP Set: Subset of 325 libraries from the Complete Set. Only non-normalized libraries derived from normal tissue samples are included. (14,397 genes, 530,618 EST

sequences).

CGAP Lung Set: Subset of 16 libraries from the Complete Set. Only non-normalized libraries derived from a normal lung tissue are included.

Lib8840: Single largest UNIGENE EST library, from normal pancreatic islet cells (4447 genes, 40,083 EST sequences) [72].

3.5.3 Identification of Alternative Splicing Events

For each gene, we compare the intron structure of the major isoform with the intron structure of each EST sequence. If an EST sequence contains at least one intron that differs from the corresponding major isoform intron at the 5' or 3' splice site, that EST is counted as an alternative transcript. The total number of alternative transcripts is defined as the total number of ESTs containing alternative splicing. The fraction of alternative transcripts is defined as the number of ESTs with alternative splicing divided by the total number of ESTs for a gene. The number of isoforms for a gene is defined as the number of unique intron patterns discovered in the EST libraries. We also defined the number of detected splicing reactions as the total number of introns observed in all EST sequences of a gene (Illustrated in Figure 3.9).

3.5.4 Microarray Based Abundance Measure

We have used microarray data from the NCBI GEO Series 2719 [78] for our study. These data cover a wide range of normal and pathogenic tissues, across many different tissue types. For this analysis only normal tissue samples were included. The comparison between microarray signal values and ESTs counts per gene in the CGAP Subset is shown in Figure 3.10. For each gene we compute average signal values across 15 normal tissue

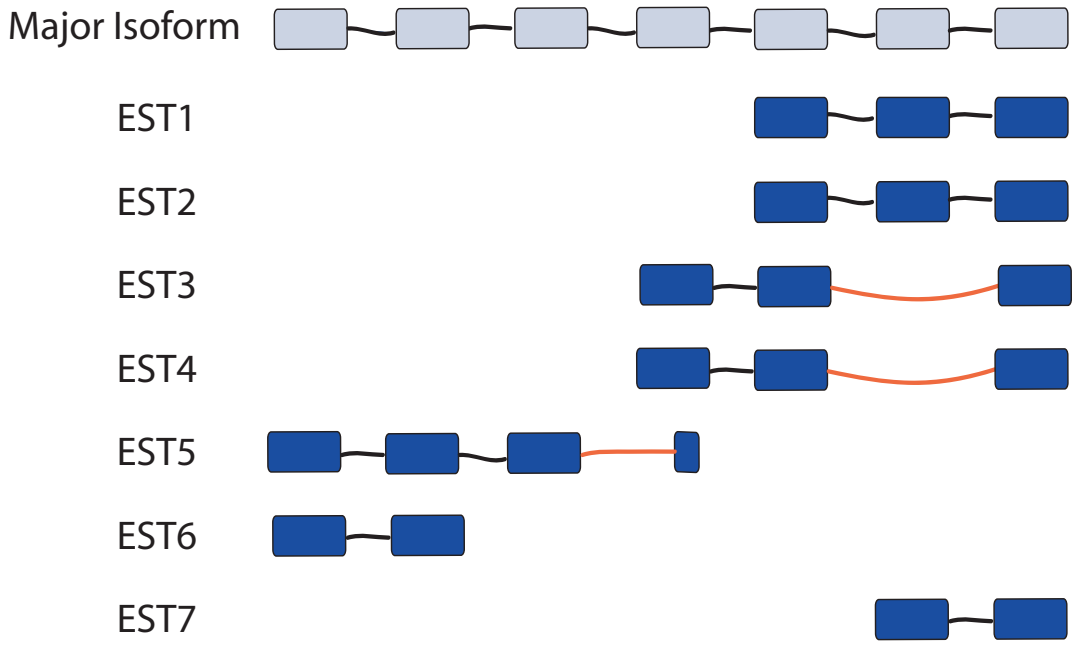

Figure 3.9: Example analysis of real EST sequences. In this hypothetical example, the major isoform of a gene has 6 introns and 7 EST have been observed in a library. Three of the EST sequences (EST3, EST4, EST5) contain alternative introns - introns that differ at the 3' and/or 5' end from corresponding intron in the major isoform. The fractional abundance of alternative transcripts is 42% (3 out of 7). The number of isoforms for this gene is 3 (major isoform, EST3 isoform, and EST5 isoforms). EST4 is not counted as an isoform because it has the same pattern as EST3. There are a total of 13 detected splicing reactions (count of all introns from all ESTs) and 3 of these splicing reactions are classified as alternative. The implied error rate for this gene is 0.23 (3 out of 13 splicing reactions).

samples from the microarray series. The genes were grouped into 100 equal size bins, based on the average signal values, and within each group the mean number of observed ESTs and the mean microarray signal were calculated. The signal value is a measure of probe intensity and has been shown [2] that log(probe intensity) is linearly proportional to log(transcripts per cell). We find a strong correlation between number of ESTs per gene and microarray signal values (correlation 0.93) on a log-log scale. Based on the fit between microarray signal and ESTs per gene, we use the following formula to estimate the number of transcripts of each gene in a cell

$$N(transcripts) = C * \alpha * X^k$$

Where X is the microarray signal value, α=0.55 and k=0.64 are values obtained from the fit of EST counts to microarray signal (Figure 3.10), and C=2 is a scaling constant to make the total number of generated transcripts equal to 800,000, approximately the content of a single Human cell [2].

3.5.5 Binary Transcript Representation

Using the intron counts, error rate, and numbers of transcripts per cell, we simulate the intron structure of a set of transcripts for each gene, as many transcripts as in a single cell. Figure 3.11 gives an illustrative example, for a gene with 6 introns and 10 transcripts. The intron structure of a given transcript is encoded as a binary string of length equal to the number of introns in the major isoform. The alternative introns - introns that differ from the major isoform in location of the 5' or 3' splice site, are represented by the symbol "1", while introns with same genomic coordinates as the major isoform are represented by the symbol "0". In this schema, transcripts 1, 3, 6 and 10 encode the major isoform of the gene, producing the string "000000". Transcripts 2, 4, and 6 contain exon skips that are

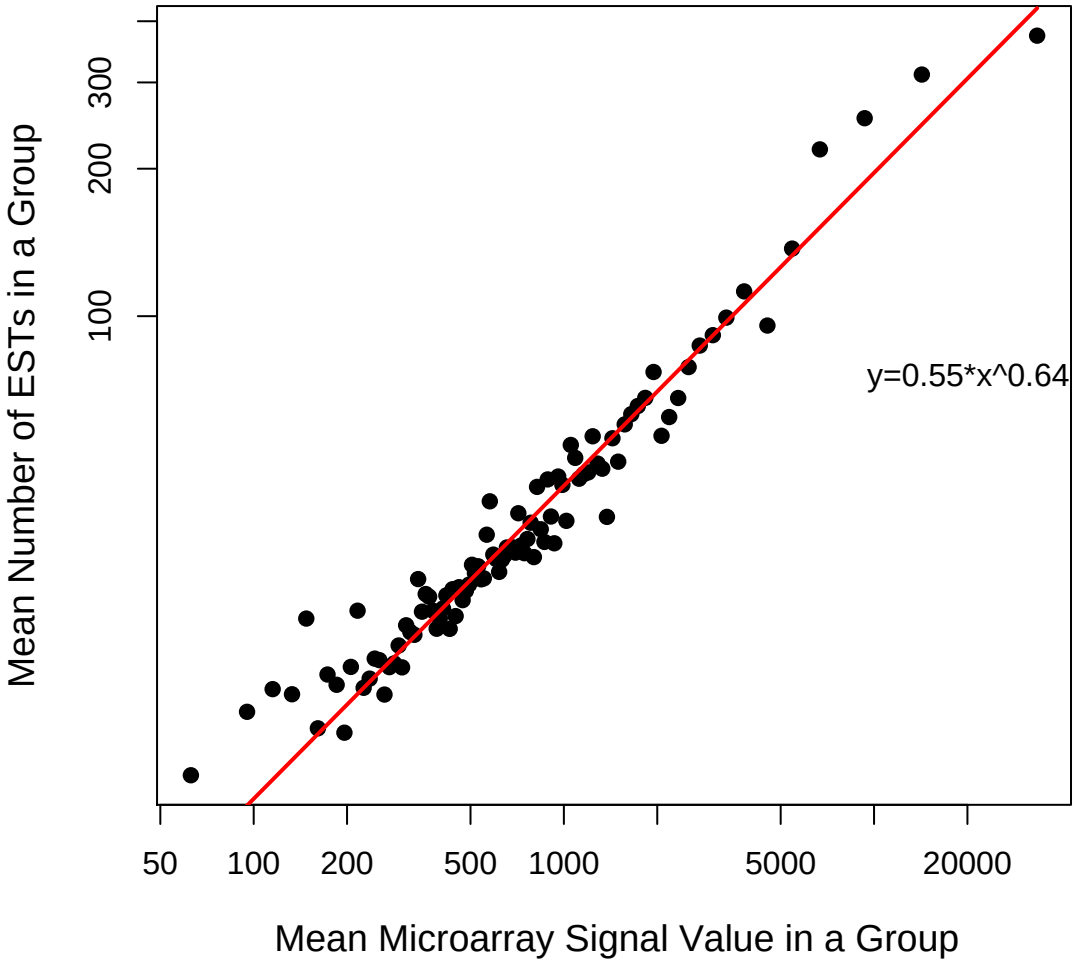

Figure 3.10: The fit between microarray signal values and number of observed ESTs. Genes from the CGAP subset were binned into 100 groups based on number of ESTs per gene. Within each bin, the mean number of ESTs and the mean microarray signal were calculated. The power law fit is shown as a red line.

different from the major isoform for two introns, thus producing ”011000” , ”000011”, and ”000011” strings. Transcripts 7, 8, and 9 contain alternative 5’ and 3’ splicing events, that modify only one intron, thus producing ”000001”, ”000001”, and 001000” strings respectively. In generating the strings, exon deletions ”11” are chosen 80% of the time, and alternative 5’ and 3’ end formation the remaining 20%, in accordance with the overall ratio found in cDNA data (see Chapter 2).

3.5.6 Simulation of Sampling

For each gene, we note the number of observed ESTs in the EST libraries, and the number of introns that are included in each EST. For each observed EST sequence for that gene, we then randomly pick one of the simulated transcripts, mimicking the clone selection step in an experimental EST protocol. Each selected transcript is then truncated, to include only those introns in the EST, simulating the partial coverage of a message typically obtained with an EST.

The sampling procedure is also illustrated in Figure 3.11. The hypothetical gene in the illustration is estimated to have 10 messages per cell, and the experimental EST libraries contain seven ESTs. Thus we randomly pick seven out of 10 simulated transcripts and truncate each intron pattern to correspond to the number of introns covered in the experimental EST sequences (highlighted in blue). The truncated patterns containing at least one ’1’ symbol represent detected alternatively spliced transcripts. For example, the full intron pattern of transcript 2 is ”011000”, but since only the 2 introns are covered in the corresponding EST sequence the pattern is truncated to ”00”, thus resulting in an undetected alternatively spliced isoform.

We obtain the number of alternative splicing transcripts for a gene by counting the number of transcripts with at least one detected alternative splicing event. We calculate

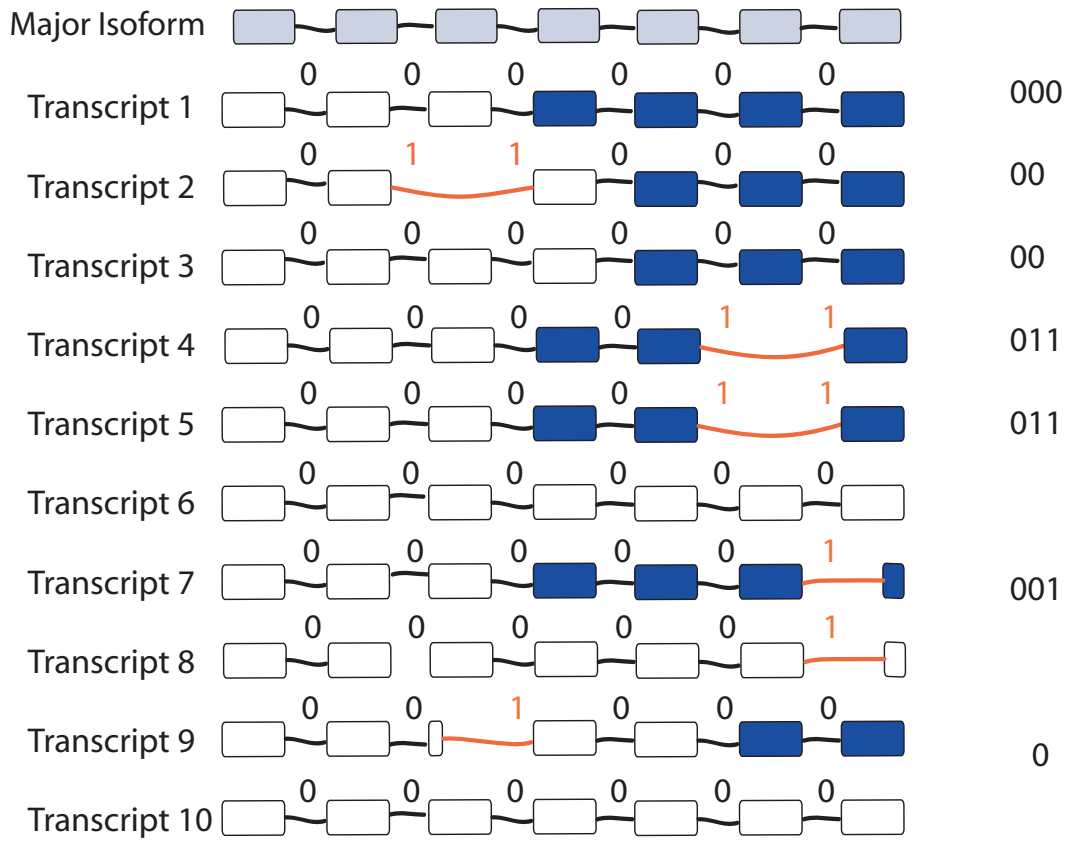

Figure 3.11: Simulation and sampling of the isoform composition of a gene with 10 messages per cell and six introns. Exons are shown as rectangles. Alternative splicing events are indicated by red intron bridges. The binary intron representation is shown above each bridge, with the symbol "1" indicating an alternative splicing event, and the symbol "0" representing a major splicing event. In the set of 10 there are total of six alternative transcripts (those with at least one '1': 2,4,5,7, 8 and 9) with four unique alternative isoforms (two patterns occur twice: 4 and 5, and 7 and 8). In this example, we assume that EST sampling extracted seven of the 10 (those with some blue exons), and that the partial message sequencing only included the colored exons. With this particular sampling, three alternative transcripts are selected (4, 5 and 6), containing two of the four unique alternative isoforms (represented by the patterns 011, and 001). Although a third alternative isoform (2) was selected, the EST sequence does not extend far enough into the message for the difference to be detected.

the number of alternative isoforms by counting number of unique splicing patterns. For example, the hypothetical gene in figure 3.11, there are a total of three detected alternative transcripts (transcripts 4, 5 and 7). The number of detected alternative isoforms for this gene is two, since transcripts 4 and 5 encode the same pattern, 011. The fraction of alternative transcripts is defined as the number of sampled alternative transcripts divided by the total number of sampled transcripts, in this case 3 out 7.

3.6 Supplementary Data

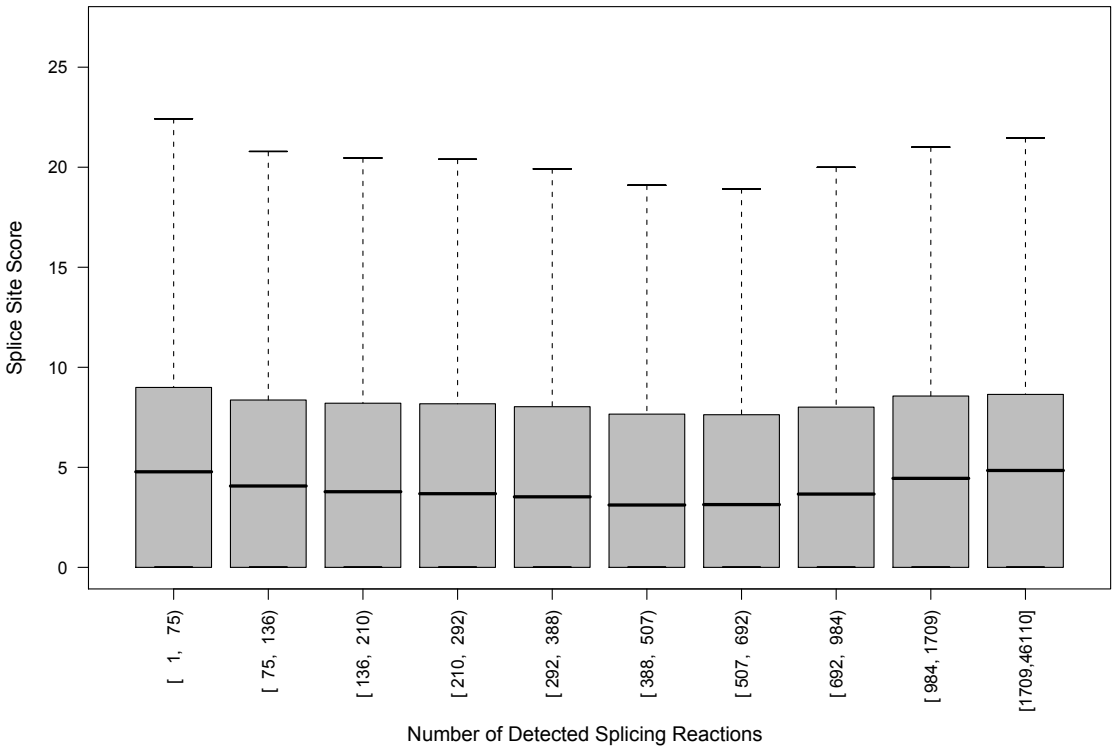

Figure 3.12: Strength of splice site motifs as computed by GeneSplicer HMM [79]. Genes from the Complete Set were divided into 10 equal size groups based on the number of detected splicing reactions per gene, and the distribution of splice site scores for the donor splice sites within each group was calculated. The scores were computed using GeneSplicer HMM program. Only scores for the splice sites present in the major isoform of a gene were used in the calculations. If splice site was not detected it was assigned a score of zero. By this measure, splice sites strength signals do not show the dependence on number of splicing reactions predicted by model 3. Similar results were obtained for acceptor splice sites (data not shown).

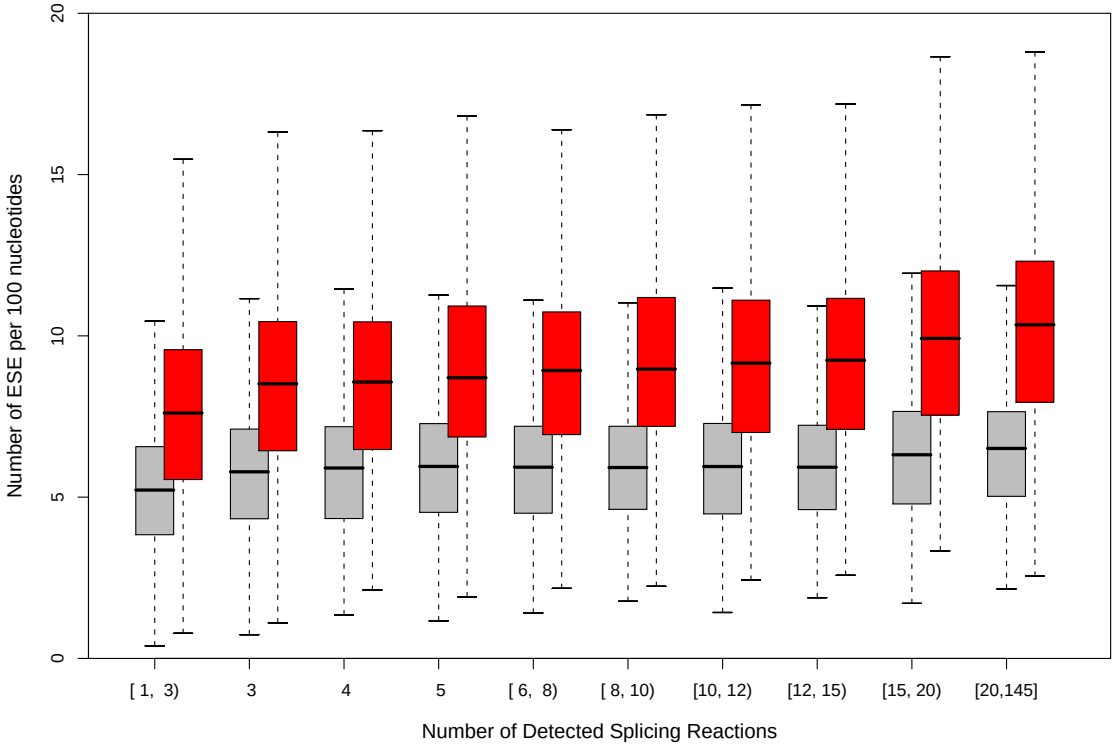

Figure 3.13: Predicted Exon Splicing Enhancers (ESE) Motifs vs. Number of Splicing Reactions. Genes from the Complete Set were divided into 10 equal size groups based on number of detected splicing reactions per gene. For each gene we calculated the number of ESE motifs present in the mRNA sequence of the major isoform, normalized by length of mRNA sequence (red bars). To make sure that signal is not due to compositional biases, we also calculate the number of ESE motifs in shuffled mRNA sequences (grey bars). As a source of ESE data, we used 238 nucleotide motifs from the RESCUE-ESE program [80]. The number of motifs rises steadily with increase in number of splicing reactions, as predicted by model 3.

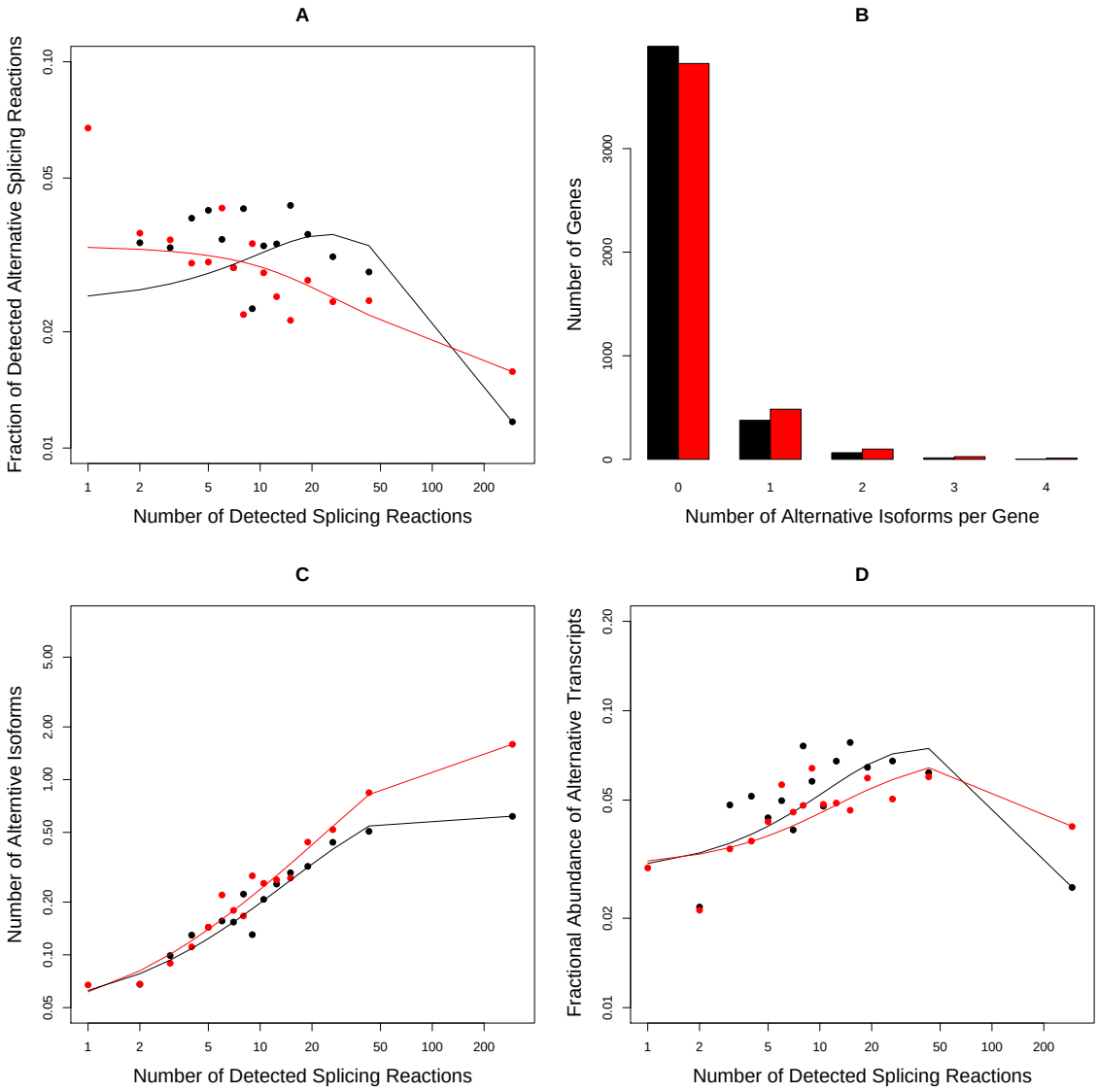

Figure 3.14: Simulation of sampling using EST data from Lib8840. Sampling in a virtual cDNA library with 10000 cells. Transcripts generated with error rate implied by Equation 2 and $\alpha = 0.4$ and $\beta = 0.015$. Simulation in red, observed data in black (Lib8440 Library, 40,083 ESTs). Panel A: Error rate as simulated by the model compared to observed implied error rate. Panel B: Number of Detected isoforms per gene distribution. Panel C: Increase in number of detected isoforms as a function of the number of detected splicing reactions. D. Fractional abundance of alternative transcripts. The results of simulation in the single EST library (Lib8840) show the same trends as simulation results in the CGAP library pool, clearly indicating that observed trends are not due to EST library pooling.

MODEL	PLOT	k=N(param)	N(obs)	RSS	AIC
Model 1	Plot A	1	50	0.00034	-396
Model 1	Plot B	1	10	22212	102
Model 1	Plot C	1	50	1.69	28
Model 1	Plot D	1	50	0.00031	-400
Model 2	Plot A	1	50	0.00038	-391
Model 2	Plot B	1	10	13050	96
Model 2	Plot C	1	50	1.39	18
Model 2	Plot D	1	50	0.00035	-394
Model 3	Plot A	2	50	0.00026	-407
Model 3	Plot B	2	10	5324	89
Model 3	Plot C	2	50	0.385	-43
Model 3	Plot D	2	50	0.00002	-411

Table 3.1: The Akaike information criterion (AIC). AIC measures a goodness of fit of the models to the observed data with a penalty for increase in the number of parameters [81] $AIC = N * log(RSS) + 2 * k$. The preferred model has a lowest AIC measure. RSS is the sum squared residuals (difference between observed data and predicted data) normalized by the number of observations. For completed description of models, plots, and parameters see Figures 3.5, 3.6, 3.7. As can be seen in the table, the lowest AIC scores are obtained by model 3.

Chapter 4

Chapter 4. Protein Stability of Alternatively Spliced Proteins

4.1 Abstract

It is frequently argued that alternative splicing is one of the principal mechanisms for generating functional diversity in higher Eukaryotes. Even though nearly every Human gene has at least one alternative splice form, very little is so far known about the structure and function of resulting protein products. In this study, we investigate the implied impact of alternative splicing on protein sequence and structure. For most proteins, structural stability is necessary for biological activity, providing a means of assessing the functional viability of the products of alternative splicing. We examine effects of alternative splicing on protein sequence and structure in three sets of alternative splicing events: alternative splicing events conserved across multiple species, alternative splicing events in genes that are strongly linked to disease, and all observed alternative splicing events. We find that alternative splicing events conserved across species tend to maintain protein structural integrity to a greater extent than is found in the full splicing set. However, these events represent only a small fraction of the total. We predict that the majority of alternative isoforms result in unstable protein conformations. Alternative splicing in disease-associated genes produces unstable structures just as frequently as all other genes, clearly indicating that selection to reduce the effects of alternative splicing on this set is not especially pronounced. We find that overall, the properties of alternative spliced proteins are consistent with the outcome of random errors made by splicing machinery.

4.2 Introduction

Bioinformatic analysis of EST sequences as well as microarray experiments shows that at least 78% all genes in Human undergo alternative splicing, producing an average of four isoforms for every gene [18, 38, 20, 21]. Three broad hypotheses have been proposed to explain the prevalence of alternative splicing in higher Eukaryotes. The first hypothesis is that alternative splicing generates functional diversity by producing alternative protein products [6, 82, 24]. The second is that alternative splicing acts as a regulation mechanism to control the level of useful gene products, by means of changing the fraction of functional and non-functional transcripts [33, 36, 66]. The third hypothesis is that alternative transcripts are noise, that is, the result of occasional mistakes made by the splicing machinery (Chapter 3)

Although there are some well known examples of alternative splicing generating protein functional diversity such as in Dscam [83], NOVA [17], Neurexin [84], and CD44 [14], the vast majority of alternative transcripts have no known function at the protein level. Alternative splicing may also be functional at the message level, acting as a mechanism for switching off or down regulating the expression of a protein [36]. Although there are some well established examples such as Drosophila sex-lethal (Sxl) [15], mdm2 [85], ABCC4 [86], MID1 [87], hUPF2 [88], here too, in nearly all cases, no such function is known. A recent microarray-based survey of NMD effects by Pan et al. has found that transcripts with premature stop codons (PTC) are present at a uniform level in Human tissues and only a small fraction of PTC transcripts are substantially regulated by nonsense mediated decay (NMD) [35]. Nevertheless, functional alternatives and regulation hypotheses are generally regarded as the most plausible explanation for the large number of alternative isoforms [89].

The noise hypothesis is diametrically opposed to the functional point of view, and

argues that most of alternative transcripts are generated as a product of spliceosomal mistakes in selection of splice sites. We have shown in Chapter 3 that this hypothesis is consistent with a number of non-trivial properties of the distributions of isoform abundance and diversity.

In this study, we investigate the impact of alternative splicing on protein sequence and structure. Using random exon deletions as a control, we examine the effect of observed deletions on structural properties such as exposed hydrophobic area, loss of contacts, and length of the gap created in the polypeptide chain. We also contrast the properties of alternative splicing events found only in Human with those conserved across multiple species and so expected to be more likely to be functional [31, 39, 90] . Splicing in monogenic disease genes is also investigated, since these are likely to be under the strongest selection pressure to maintain function.

We find that on average, alternative splicing events have a markedly deleterious effect on protein structure, similar to that found for random exon deletions, and so are unlikely to encode for alternative protein function. Splice forms conserved across multiple species on average have a less severe impact on structure, although in many cases are still very disruptive. We find that disease associated genes do not show special sensitivity to alternative splicing, clearly indicating that there is no strong selection to remove deleterious changes introduced by alternative splicing. Overall, our prediction is that the majority of alternative proteins are structurally unstable and if expressed will be without function, consistent with the noisy splicing hypothesis.

4.3 Results

4.3.1 Difference in Protein Sequences of Isoforms

We first examine the effect of alternative splicing at the protein sequence level. We compiled an initial set of 85,136 isoforms from 19,743 genes using sequence data obtained from the Refseq [91], Unigene [92], and Hinv [58] databases. All isoforms in this set were subjected to quality control checks to ensure that every splice junction is valid. An additional 18,995 isoforms were removed because of uncertainty in the alignment at the 5' or 3' ends. Transcripts from genes with no observed alternative isoforms were also eliminated, leaving 10,972 genes with two or more isoforms each. There are a total of 55,217 isoforms, of which 40% (20,998) were derived from cDNA sequence, and the remaining 60% (34,219) were derived from EST sequences. Partial isoforms derived from ESTs were completed by copying missing exon structure from the corresponding major isoform (see Methods).

All validated isoforms were translated to provide the corresponding protein sequences, allowing for possible errors in N terminal position (see Methods). For each gene, one of the cDNA derived isoforms was selected as the major isoform. To obtain an overall measure of protein sequence length differences between minor and major isoforms, we subtracted the length of a major isoform from that of each minor isoform of the same gene and plotted the histogram of length differences (Figure 4.1). Approximately 20% of all minor isoforms had the same protein sequence length as the major isoform, mostly due to alternative splicing outside the coding regions (not shown in the plot). Most minor isoforms are significantly shorter than major isoforms ($\sim$70%) and only 9% of all minor isoform are longer than major isoform. The signal is dominated by protein length differences of more than 100 amino acids ($\sim$43%).

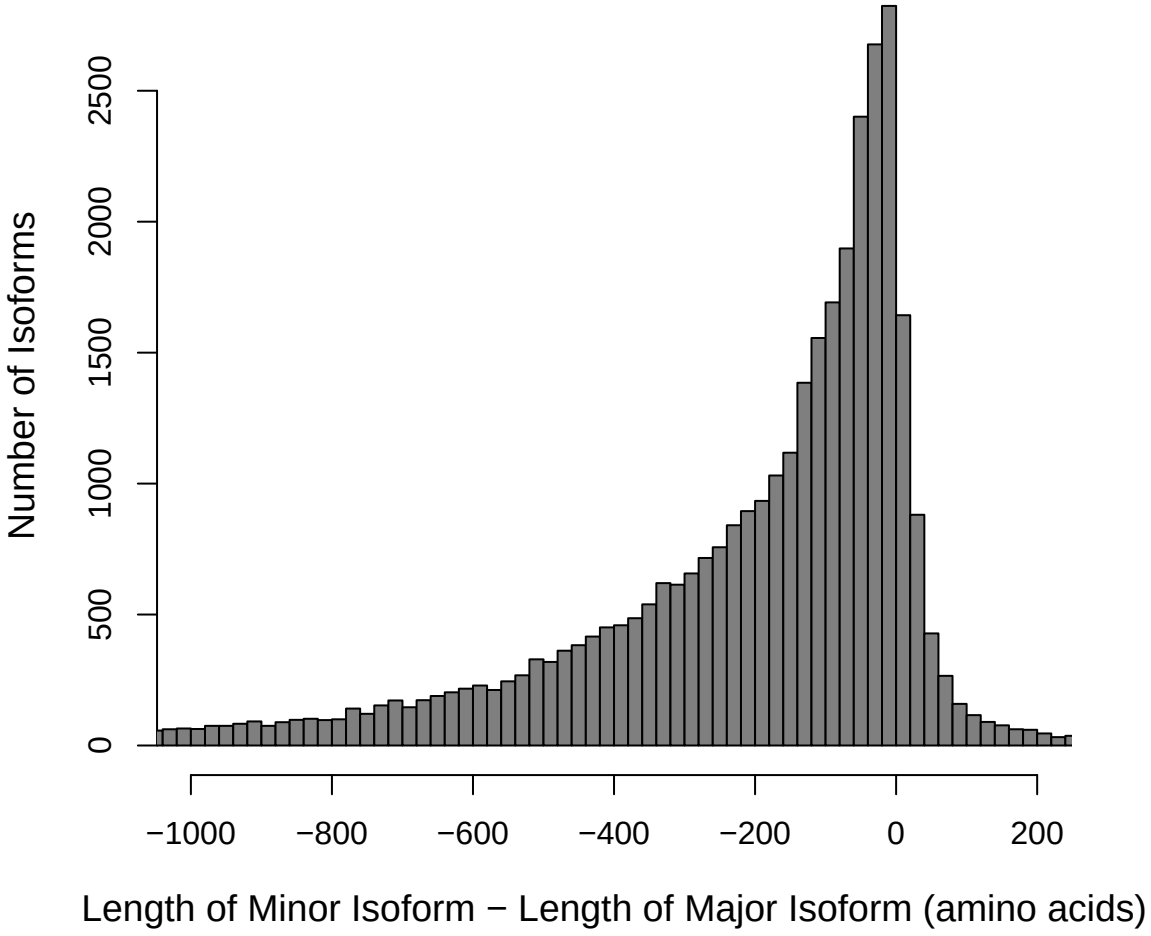

Figure 4.1: Distribution of differences in amino acid sequence length between proteins coded for by major and minor isoforms. Most minor isoforms are substantially shorter than the major isoform.

We repeated this analysis on the subset of isoforms derived only from full length cDNA sequences to make sure that these observations are not due to incorrect reconstruction of messages from partial EST sequences We found that both datasets have approximately the same distribution of length diff erences. The cDNA dataset has an increased fraction of longer isoforms ($\sim$17% rather than $\sim$9%) and a correspondingly smaller fraction of same length isoforms, but the fraction of shorter major isoforms remains the same in both sets. 10,160 isoforms are predicted to be subject to the NMD degradation pathway (prediction based on 50 nucleotide rule) [33]. Removal of these results in a reduction in the fraction of isoforms more than 100 amino acids shorter than the major one (43% to 36%). The same effect is observed in the cDNA only subset.

In order to analyze the differences between isoforms further, we aligned the implied protein translation of each minor isoform to that of the corresponding major isoform. The alignment is preformed in two steps, first we align the major and minor isoforms' exons to a common genome reference frame. We then use the cDNA alignment to create aligned protein translations. Figure 4.2 illustrates the procedure.

We term each continuous stretch of difference within an alignment between two isoforms a protein splicing fragment (PSF). In cases where alternative splicing lies outside the coding region, the protein alignment will be identical and thus no PSF will be produced. On the other hand, if there are multiple alternative splicing events within a coding region, multiple PSFs will be generated from a single major/minor comparison. For example, in the TPM2 (tropomyosin 2) gene (Figure 4.3), there are two alternative splicing differences between the major and minor isoforms, generating two PSFs.

The protein splicing fragments were classified into three broad categories: deletion, insertion, and substitution. Substitutions were further classified into three classes: perfect replacement - a fragment is replaced with another fragment of the same size; truncation - a fragment is replaced with a smaller fragment; and elongation - a fragment replaced

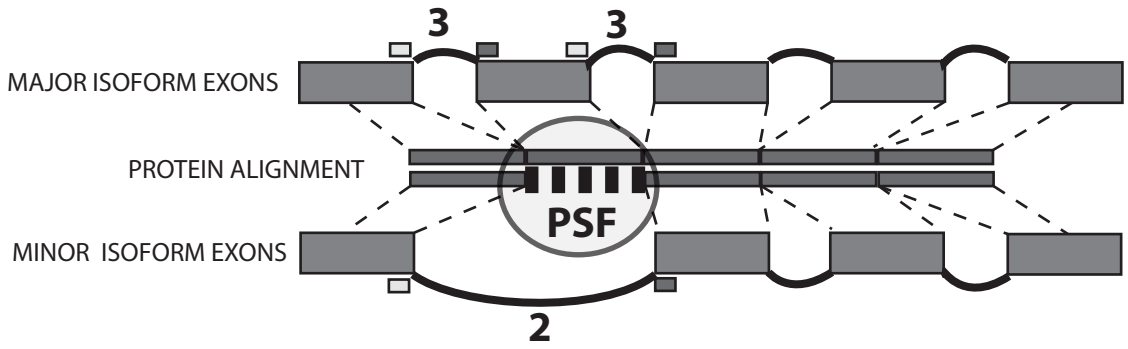

Figure 4.2: Identification of protein splicing fragments (PSFs). Protein sequences corresponding to the major and a minor isoform of a gene are aligned, and regions in the alignment that differ identified. Deletions are defined as missing fragments in minor isoforms. Replacement is defined as a substitution of identical size. Truncations and Elongations are substitutions that change the length of a fragment. Numbers above each intron bridge are conservation scores, the number of species in which this or a homologous bridge is found. Here, the alternative intron has a score of 2, indicating it was detected in Human and one other species (maximum conservations score is 11)

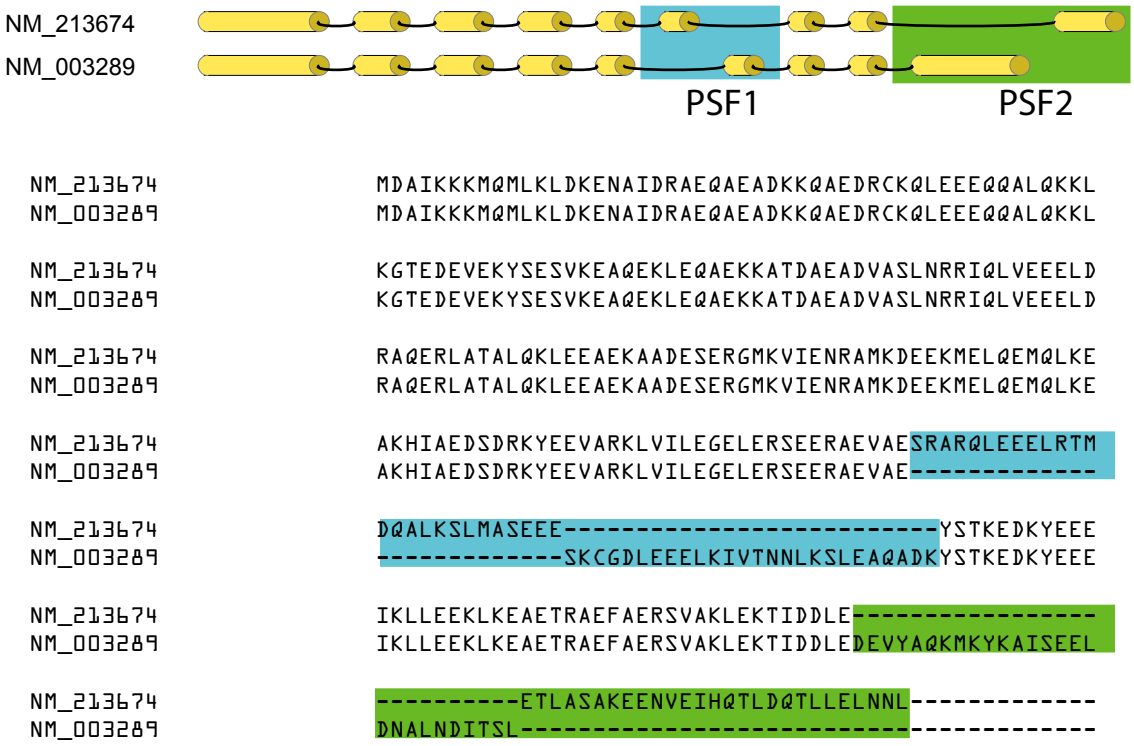

Figure 4.3: Example of Protein Splicing Fragments (PSFs) in tropomyosin 2 (TPM2). The exon alignment of two isoforms shows two PSFs. In this case, both are the result of exon swaps (an exon in one isoform is replaced by a different one in the other isoform), and in each instance, the replacement exon is the same length as the major isoform one. The PSF in the middle of the sequence is thus classified as an internal replacement, and the one at the 3' end is a C-terminal Replacement.

with a larger fragment. We also classified PSFs into four broad classes based on location: N-terminal, Internal, C-terminal and Not Classified. The last category is reserved for differences that extend over the entire length of the alignment. As an example, consider the alignment between two isoforms of tropomyosin 2 (TPM2) gene in Figure 4.3, where there are two replacements, one creating an internal 26 residue long substitution, the other also a 26 residue replacement, at the C terminus.

	Not Classified	Cterm	Internal	Nterm	Fraction (%)
Truncation	3000	9425	1252	2336	47.2
Deletion		278	4603	8359	39.0
Elongation	369	1070	507	537	7.3
Insertion		25	1597	444	6.1
Replacement	1	39	68	35	0.4
Fraction(%)	9.9	31.9	23.6	34.5	100

Table 4.1: Classification of Protein Splicing Fragments (PSFs). Subsequences affected by splicing are classified by the effect on length (Replacement - same length fragment, Truncation - shorter fragment than in the major isoform, Insertion - longer fragment, Deletion, Insertion) and by location in the open reading frame (Internal, C-terminal, N-terminal, and Not-Classified) The majority of PSFs belong to the Deletion and Truncation categories.

The results of the protein splicing fragment classification are summarized in Table 4.1. As already implied by the whole protein length comparisons, the majority of fragments are Deletions and Truncations (87.6% of all fragments). Many of the C-terminal Truncations are produced as a result of premature stop codons due to frame shifts. Many of these are predicted to be degraded by NMD by the ”50 nucleotide rule” [93], although this prediction might not be correct: Recent evidence by Pan et al, suggests that significant fraction of transcripts with premature stop codon will not be affected by NMD [35]. The most common types of deletions are N-terminal deletions. These are largely generated through alternative promoters, and strictly speaking, should be considered separately from other alternative splicing events, since different machinery is involved. As far as is known, there is no quality control mechanism similar to NMD for deletions on the N-terminal ends of proteins, thus we would expect that these proteins are actually produced.

Replacement of a protein fragment with another protein fragment of exactly the same size is the least common type of fragment (0.4%). Although rare, these events are highly expressed (supported by large number of EST observations) and we will show in the next section that these events exhibit the strongest conservation signal across multiple species.

4.3.2 Conserved Alternative Splicing Subsets

To obtain a splicing subset enriched for function, we compiled a list of alternative splicing events conserved across multiple species. Although it is reasonable to assume that cross species conservation is correlated to functionality, there are two caveats to be borne in mind. First, the conserved subset is biased toward genes expressed in high abundance in other species, since they are more likely to be detected in EST experiments. Second, high abundance genes are also likely to produce more alternative isoforms as a result of noise (Chapter 3), so some of the isoforms are likely to be non-functional. Nevertheless, several evolutionary trends have been shown to be correlated to conservation of alternative splicing events, such as increase in selection pressure against synonymous mutations and an increase in selection pressure for protein reading-frame preservation (reviewed by Xing et al. [32]), supporting the idea of at least enrichment for function in conserved events.

To find conserved alternative splicing events, we searched sequences of all exon-exon junctions for hits in transcripts of other species, and defined the conservation score for each junction as the number of species that had at least one significant hit to that junction. We then define the conservation score for each protein splicing fragment (PSF) as the maximum conservation score from all of alternative splice junctions that underlie that PSF. The distribution of PSF conservation scores from all isoforms except those predicted to be subject to NMD is shown in Table 4.2.

	Human Only	2 Species	3 Species	4+ Species
Replacement	0.4	0.6	6.2	9.8
Deletion	50.6	38.7	27.6	29.4
Truncation	34.2	29.6	29.2	21.6
Elongation	7.7	15.7	26.9	33.3
Insertion	7.0	15.4	10.0	5.9
Total	100%	100%	100%	100%
N(PSFs)	23287	1761	438	51

Table 4.2: Cross-species conservation of Protein Splicing Fragments (PSFs). Splicing effects most likely to be functional (replacement, and elongation and insertion) are markedly enhanced in the conserved sets.

The perfect length replacements show strongest conservation. They represent 9.8% of all PSFs conserved across four or more species - a 25 fold increase from the 0.4% value in the Human only subset. The fraction of Insertions and Elongations also increases with increasing conservation, while Deletions and Truncations decrease. The most obvious explanation for these observations is that deletions and truncations have a greater tendency to disrupt protein structure and are thus less likely to be conserved. As we observed in the distribution of length changes (Figure 4.1), deletions and truncations tend to remove a large numbers of residues, typically more than a 100. Of course, perfect replacements are least likely to effect protein structure, since they preserve protein length. Insertions and Elongations tend to change the length by fewer than 25 residues, and are thus also less likely to disrupt structure. This effect is investigated further in the next section.

The relationship between change in length and conservation across species is shown in Figure 4.4A. Figure 4.4B shows the relationship between length change and minor isoform abundance. The assumption here is that more abundant isoforms are more likely to be functional, providing another means of examining the relationship between structure properties and function. Abundance is defined as the number of EST observations for alternative splice junctions underlying the PSF. In cases where there are multiple alternative splice junctions underlying a PSF, we use the junction with the highest EST count as

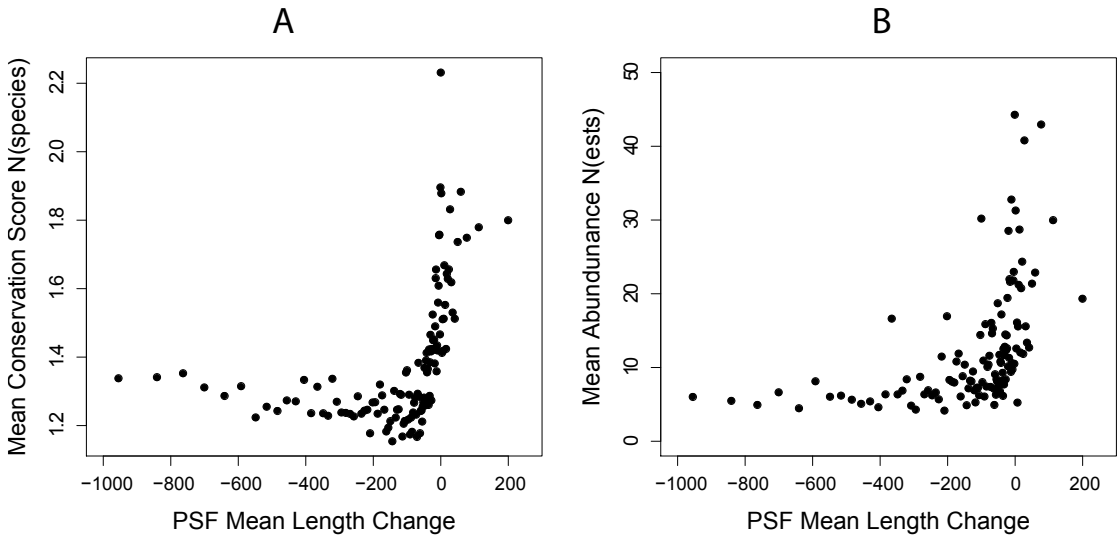

Figure 4.4: Cross-species Conservation and Abundance vs. Protein Splicing Fragment (PSF) length change. PSFs were divided into 120 groups, with at least 200 PSFs within each group, and mean conservation scores and abundance were calculated. Large changes in length are typically only found in unconserved splice forms and at low abundance.

the measure of abundance. This is an approximate measure of abundance because it does not take into account EST sampling biases, however it is expected to be proportional to actual number of copies of the isoform in the sample. The Figure 4.4 clearly shows that change in length is highly correlated to both abundance and conservation across species. Small changes are both conserved and abundant, suggesting that they are more likely to be functional.

4.3.3 Properties of Disease Gene Subsets

We have also analyzed the frequency of alternative splicing events in a subset of genes strongly associated with Human disease. Disease associated genes were obtained from three databases: OMIM [94] (2241 genes), HGMD [95] (834 genes), and Genetests [96] (1000 genes). We also compiled a CORE set of 530 genes found in all three databases. Although the exact mechanism leading to disease is different in each

case, in general the cause can be attributed to reduced total activity of a protein product as a result of a mutation of some kind.

The extent of change to protein sequence as a result of alternative splicing is far greater than change due to a typical single amino acid mutation, and as a consequence the likelihood of disruption of function is also greater. But unlike a deleterious amino acid mutation, which will affect all transcripts of a gene, alternative splicing only affects a fraction of all transcripts. As long as the fraction of alternative transcripts does not exceed some level where the normal function of a gene is seriously affected, there will be no pressure to reduce deleterious changes. Assuming that these assumptions are correct, disease genes should undergo alternative splicing with the same frequency as all other genes.

Figure 4.5 shows the distribution of the number of isoforms per gene for all Human genes in our database (19,743 genes, 85,136 isoforms) and for the various subsets of disease associated ones. Except for the first set of bars (single isoform, i.e. no alternative spicing), all the sets show nearly identical distributions. Abundance of alternative isoforms (as measured by fraction of all transcripts per gene that are alternative) is nearly identical (8% all vs $\sim$ 7% disease). We also looked at the distribution of overall length change between major and minor forms of proteins in disease associated genes, which also shows no significant differences between gene sets. We did not find significant differences in predicted NMD fraction, or the types and locations of PSFs. Based on these observations, we conclude that pressure to reduce the frequency, or severity of impact of alternative splicing events is the same for disease genes as non-disease genes.

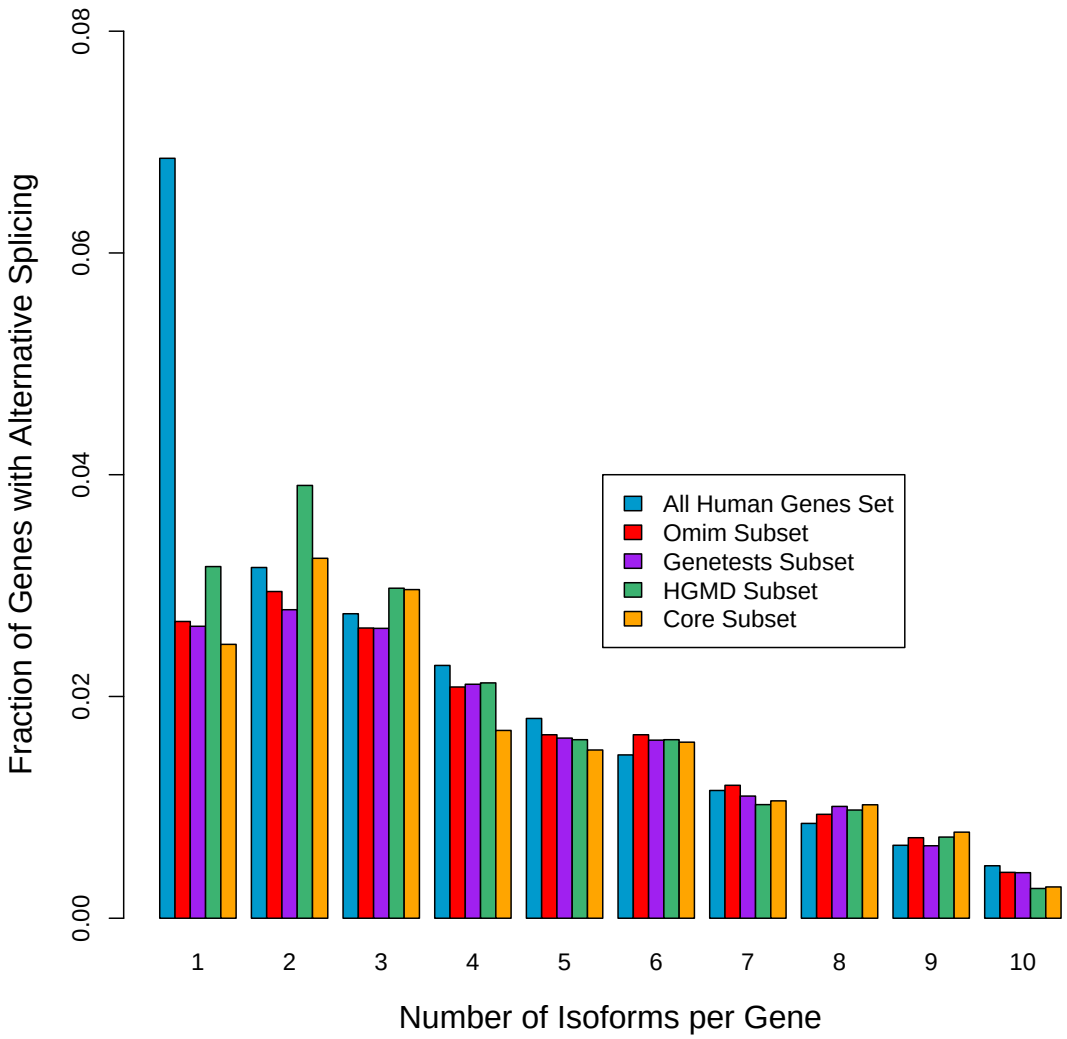

Figure 4.5: Fraction of genes with alternative splicing in disease associated genes. All genes (19,743 genes, 85,136 isoforms), and three subsets of disease related genes derived from OMIM (2241 genes), Genetests (1000 genes) and HGMD (834 genes). The Core set are genes present in all three databases (530 genes). All sets show a similar distribution of number of isoforms per gene, with the exception of the first set of bars, which represent the fraction of genes without alternative splicing (7% vs $\sim$ 2-3% in disease subsets).

4.3.4 Stability of Protein Structures Produced by Alternative Splicing

For proteins with known or modelable three-dimensional structure, we can ask what fraction of alternative splicing events is likely to result in a stably folded protein product. For this purpose possible templates were identified in the PDB for the major isoform of each gene, the exons were mapped on to the 3D structural coordinates, and regions corresponding to protein structure fragments arising from Internal Deletions were removed (see Methods). The analysis was performed on the resulting set of modified protein structures. The impact of a deletion is measured in terms of the distance between the resulting chain ends, newly exposed hydrophobic area, total newly exposed area, and number of residue-residue contacts lost. An example of mapping of alternative splicing to structural fragments in growth hormone 1 gene (GH1) is illustrated in Figure 4.6.

There are no tools for accurate prediction of protein stability. However, deletion of a randomly chosen exon is very unlikely to result in a stable protein structure. We make use of this feature to generate a reference set of unstable structures, and compare properties of proteins produced by observed alternative splicing deletions with these. About 60% of all possible internal exon deletions result in a frame shift, and almost all of these are predicted to be degraded by nonsense mediated decay, thus they were not considered in our calculations.

The remaining 40% of in-frame deletions form a pool from which the reference set were selected. For each observed exon deletion included in the analysis, a random exon deletion with the same number of residues was found, generating a reference set with the same length distribution as the observed data. Just as with real deletions, random exon deletions were mapped to 3D structure coordinates, and regions of chain corresponding to the exon was removed. The final data sets consist of 1439 random deletions and 1439 real deletions (1085 splicing events observed only in Human, 263 in two species, and 76

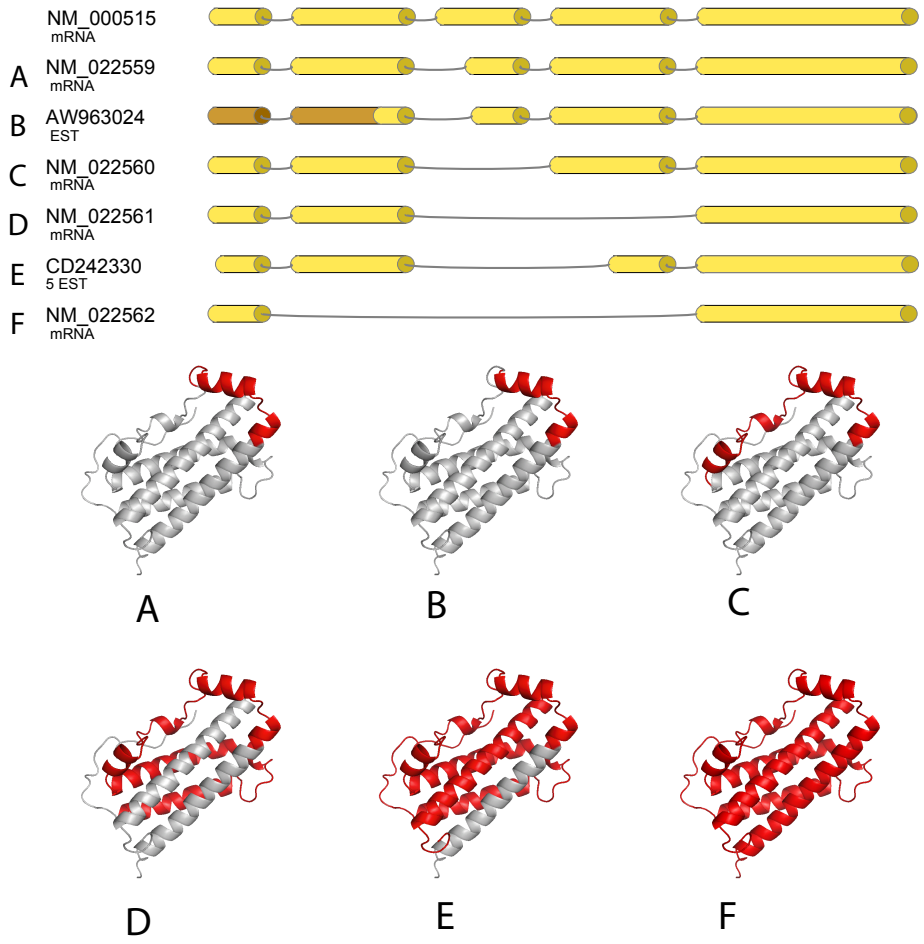

Figure 4.6: Alternative splicing in growth hormone 1 (GH1). The exon structure of the major isoform and six minor isoforms are shown as yellow bars. The location of deletions and truncations in the protein structure relative to the major isoform are highlighted in red. Isoforms A, B, C and D are classified as Internal Deletions. Isoforms E and F produce a frame shift and are classified as C-terminal Truncations. Isoform B was derived from EST sequence, and part of the exon structure (colored brown) was copied from the major isoform. The isoforms are sorted by severity of impact on structure.

in three species, and 15 in four or more species).

The distributions of the various structural features are shown in Figure 4.7. The random and full Human only sets have very similar distributions for all structural properties. We conclude from this that the large majority of alternatively spliced deletions result in the production of unstable protein folds. On the other hand, it is immediately obvious that deletions that are conserved across multiple species tend to remove fewer residues, have a smaller end-to-end distance, lose fewer contacts, and expose less total and hydrophobic surface area. That is, conserved deletions also tend to be conservative in term of structural impact, supporting the view that these sets are enriched for function compared with unconserved events.

4.4 Discussion

Alternative splicing can generate a large number of isoforms starting from a single premessege mRNA. A well known example of production of molecular diversity by alternative splicing is the Drosophila Dscam gene, which can potentially generate as many as 38,000 isoforms [83]. In Human, nearly every gene (93% , Figure 4.5) is alternatively spliced, (median 3 isoforms per gene). Most ($>$70%) isoforms change protein coding regions, and therefore potentially produce novel protein products [97]. On this basis, it is frequently argued that alternative splicing provides a mechanism for complex organisms such as Human to generate a large number of novel molecular components from a relatively small number of genes [23].

The basic assumption in this view is that products of alternative splicing are functional. However, little is known about the protein sequences and resulting protein structure of alternative isoforms. In an effort to decipher the functionality of isoforms by other means, numerous bioinformatics studies have analyzed various properties. It has

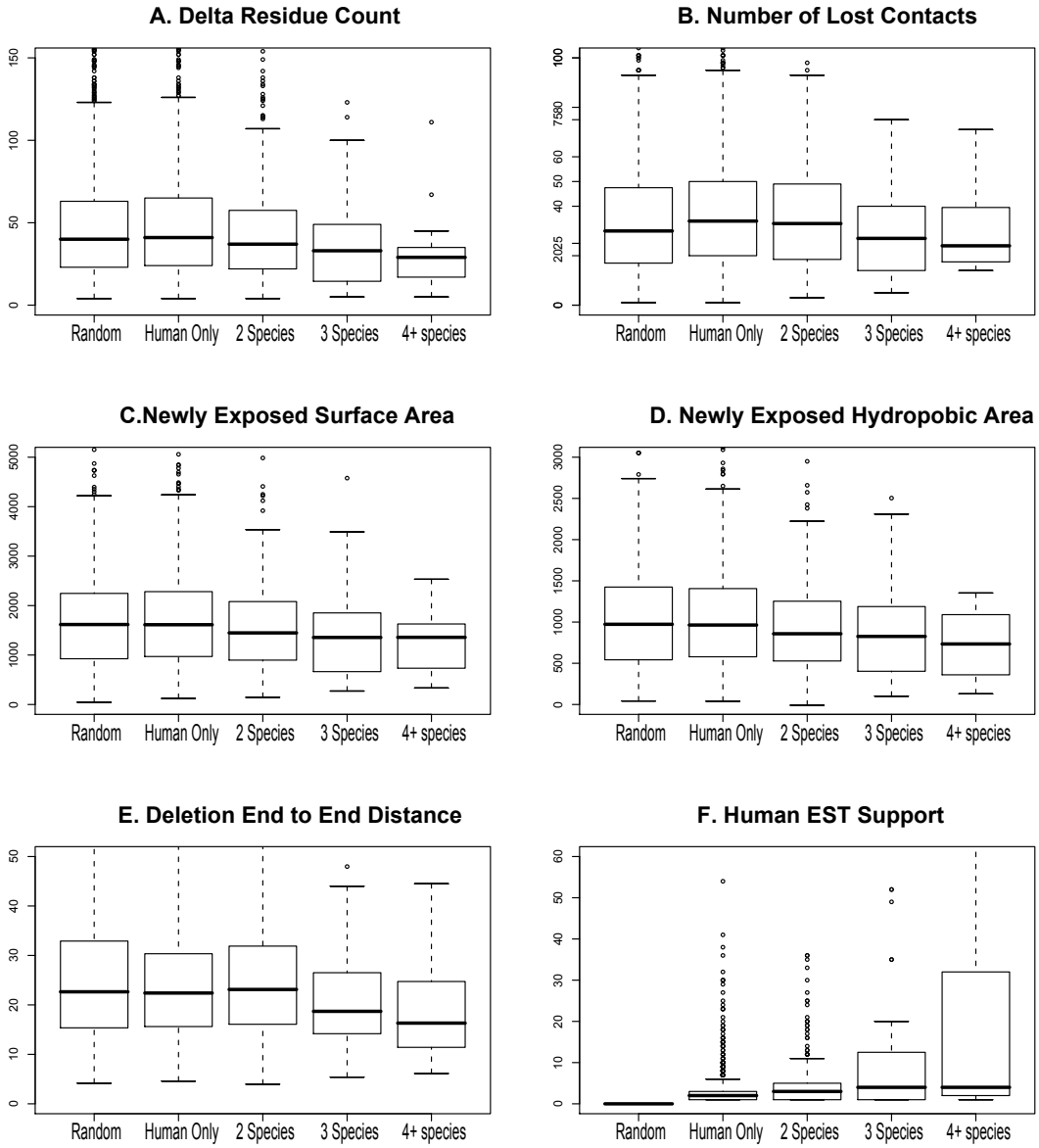

Figure 4.7: Comparison between random exon deletions and deletions observed in minor isoforms. Genes are divided into 5 sets: Random Exon Deletions, Human Only Deletions, Conserved across 2 species, 3 species, and 4 or more Species. A: Number of residues deleted (amino acids). B: Number of contacts lost. C: Newly exposed surface area (square $\mathring{A}$) D: Newly exposed hydrophobic surface area (square $\mathring{A}$) E: Distance between C-alpha atoms and ends of a deletion ($\mathring{A}^2$) F: Number of EST sequences that support the alternative splicing). For all the structural properties, the random and all Human distributions are very similar, whereas minor isoforms found in multiple species exhibit more conservative structural changes.

been found that a small fraction of alternative splicing shows clear signs of functionality, particularly those exhibiting tissue specificity [28], expression in high abundance [64], and cross species conservation [27]. Conserved isoforms especially seem tend to preserve protein coding frames and are less likely to be subject to nonsense mediated decay [39, 31, 90]. Compared to a large body of literature on the effects of alternative splicing on a sequence level, relatively little is known about its affect on protein structure and function.

Impact of alternative splicing on protein domains has been analyzed by a number of researchers [60, 98, 99, 90, 97]. Kriventseva et al. have found that alternative splicing removes/inserts whole domains more frequently than expected by chance (approximately 3 fold enrichment), while partial disruption of domains occurs less frequently than expected by chance (approximately 1.6 fold reduction) [60]. More recently, Yeo et al. also analyzed alternative spliced exons that partially overlap with INTERPRO-annotated protein domains and found that frequency of alternatively spliced exons that disrupt domain regions is lower than expected by chance, although the reduction was smaller than in Kriventseva et al. study (approximately 1.3 fold reduction) [90]. Analysis of the types of domains affected by alternative splicing indicates that many are involved in signal processing, development, and differentiation [98]. Resch et al. have found that disruption of domains involved in protein-protein interactions is the most common type of domain impact, a finding that is consistent with a regulatory function of alternative splicing [99].

There have been a few studies of the impact of alternative splicing on the secondary and tertiary structure of proteins. Homma et al. have analyzed the location of alternative splice sites and relation to SCOP domain boundaries [100]. They found that alternative splice sites occur inside SCOP domains at the expected frequency (13.2% observed vs 14.3% expected), although there is greater tendency for them to occur near SCOP domain boundaries (9.4% observed vs 3.9% expected). Furthermore, they examined relative

abundance of the variants, and found that the variants encoding unstable protein products tend to be species specific and are expressed at a significantly lower levels compared to the stable variants. Wang et al. have examined the secondary structure and surface exposure of alternatively spliced regions in alternative isoforms found in the SWISS-PROT database [101]. They found that the boundaries of alternative splicing tend to occur in coil regions and at surface exposed regions of proteins at frequencies greater than expected by chance. Wang et al. also examined 3D models of alternative isoforms via threading and molecular dynamics, and concluded that isoforms are capable of producing proteins with stable conformations. More recently, Romero et al. found that a significant fraction of alternatively spliced regions (81% of 75 alternative spliced fragments) were predicted to be disordered [84]. A proposed explanation for this observation is that alternative splicing enables functional and regulatory diversity, while avoiding structural complications by generating disordered regions in proteins (regions that lack an equilibrium 3D structure under physiological conditions).

It is clear that impact of alternative splicing on protein structure, stability, and function remains poorly understood. Although, there seems to be a general agreement that alternative splicing events conserved across species probably results in stable and functional protein products, there seems to be no consensus of species-specific isoforms.In Chapter 3 we argued that a large fraction of isoforms are products of occasional splicing mistakes in selection of splice sites. That hypothesis is supported by observations that the increase in number of isoforms is proportional to expression level and number of introns in a genes; that most isoforms are expressed at low abundance levels; and that few show clear tissue specificity [25, 34]. The main principle of the noise hypothesis is that large error rates can be tolerated as long as adequate levels of functional product are produced and toxic effects on the system are avoided. If these requirements are satisfied, there will be no selection pressure to reduce frequency of alternative splicing, and a great diversity of isoforms can be generated.

At the sequence level, we find that many alternative isoforms are predicted to produce proteins that are significantly smaller than the corresponding major isoforms. Removing isoforms that are predicted to be subject to nonsense-mediated decay does not change this outcome. These finding are in a qualitative agreement with other studies of impact of alternative splicing on protein sequences [101]. Wang et al. have analyzed alternative isoforms annotated in the SWISS-PROT database and found that deletions account for 57% of all annotated events, while insertions only account for 5% of all annoted splicing events.

Using conservation as a proxy for functionality, we find that small changes in sequence length are more likely to be conserved. Replacements that do not change the protein length show the strongest conservation signal. These observations make sense, since the smaller the change, the less likely it is to be disruptive to protein structure, increasing the likelihood of maintaining function or possibly generating new function.At the level of three-dimensional structure, we compared the impact of in-frame deletions introduced by alternative splicing to that of randomly selected in-frame exon deletions. Random deletions are unlikely to result in a stable protein fold, and so provide a reference set for testing the viability of deletions observed in real isoforms. We find that isoforms observed only in Human show the same distribution as the random ones, for all structural parameters. Deletions that are conserved across multiple species tend to be more structurally conservative - the distances between ends of deletions tend to be smaller, they expose less hydrophobic surface, and lose fewer contacts. From this observation we conclude that most species specific isoforms are unlikely to result in stable conformations.

Our analysis of disease genes did not reveal any surprising results. If alternative splicing had a negative impact on the normal functions of these genes, we should have observed strong selection pleasure to reduce the frequency and severity of such events since in this set, protein function is tightly coupled to fitness. No such pressure was ob-

served. The distributions of number of alternative isoforms, fractional abundance of alternative transcripts, number of NMD isoforms, and protein length changes in disease genes were nearly identical to those for all genes. If any pressure exists to reduce frequency of alternative splicing, it is not particularly pronounced in this set of genes.

In conclusion, our findings support the hypothesis that a large fraction of species specific isoforms are products of occasional mistakes by splicing machinery.

4.5 Methods

4.5.1 Overview

In this study, both DNA and EST sequences were collected from the Unigene, Refseq, and H-invDB databases. Sequences were aligned to Human Genome, and checked for errors. Major isoforms were selected on the basis of the most commonly observed isoform in EST libraries. See Common Methods Appendix for full description.

Briefly, we determine isoforms present in EST and cDNA data, predict protein products, compare major and minor isoforms to determine the difference in a primary sequence of the respective proteins, identify 3D templates, map sequence to structure, and compute various statistics. Each step is described in more detail below.

4.5.2 Reconstruction of Exon Structure for EST sequences

The full exon structure of EST derived isoforms must be predicted before these isoforms can be translated. We make the assumption that missing exon structure is identical to the exon structure observed in the mRNA sequence of the major isoform. On this basis,

the major isoforms are used as templates to expand each EST to a full-length transcript. For 3' EST sequences, we copy exon structure starting at the 5' end of the major isoform until we find an overlap with an exon in the EST sequence. Since EST sequences typically end at an arbitrary location within an exon, the overlapping exon is discarded, and replaced with the corresponding exon from the major isoform. An equivalent process is used to extend 5' EST sequences. For internal EST sequences, the exon structure is copied from both ends of the major isoform.

4.5.3 Location of Translation Initiation Site

Although typically the first "AUG" in an mRNA sequence is used to initiate translation, this not always correct [102], and the exact location of protein translation initiation is not generally known. To make sure that our protein translations are plausible, we find the longest translation with the translation initiation site supported by other species. This is done by first finding all possible translations of an isoform. The translations are sorted according to distance to the 5' end of the transcript, and the first 20 amino acids of each implied translation are searched against a database of all N-terminal 20 amino acid sequences compiled from 40 Eukaryotic Refseq genomes (540,000 sequences). The translation initiation sites are sorted based on the number of hits to other species, and the one with the most hits is selected.

Approximately 55% of all isoforms had translations that could be confirmed in more than two species, the remaining 45% were found only in Human. As a check of this procedure, we compared our translations of Refseq sequences with Refseq annotated translations and found 95% agreement between the two sets. In some cases single mRNA transcript can produce variety protein sequence through leaky translation [103]. In our analysis, we assume that each isoform produces a single protein sequence.

4.5.4 Protein Splicing Fragments (PSFs)

Protein translations of isoforms were aligned using the genomic coordinates of the underlying exons. This is done by first generating an mRNA alignment between major and minor isoforms pairs to genomic sequence, and then using the mRNA alignment to generate protein alignment. Protein splicing fragments (PSFs) are defined as regions within the protein alignment that differ. The alternative splicing event(s) that are responsible for producing differences in protein sequence are identified by looking for alternative introns in the region underlying each PSF. (Figure 4.2).

4.5.5 Conservation of Splice Junctions

We search a 40-nucleotide sequence around each splice junction (20 into each exon) against all EST and mRNA sequences of all homologous genes from other species. Gene homology information was obtained from the NBCI HomoloGene [70] database. Transcripts of homologous genes were obtained from the UNIGENE database. All 40 nucleotides must align with a minimum E-score of 0.01 and no more than 2 gaps to the corresponding fragment in the homologous transcript. We define the conservation score for each junction as the number of other species that had at least one significant hit to that junction. For example, if a junction was detected in mouse, rat and Human, it would receive a conservation score of 3. This procedure was repeated for all exon-exon junctions from all isoforms in all genes.

4.5.6 Conservation Score for PSFs

Using the mapping between exon structure and protein translation, we find the subset of introns in the major and minor isoforms underlying each PSF. By comparing ge-

nomic coordinates of minor and major introns within each PSF's intron subset, we identify all alternative intron pairs responsible for production of the PSF. The conservation score for a PSF is taken to be the maximum species conservation score of all the minor isoform introns. For example, if an exon insertion event resulted in an Insertion PSF with two alternative splice junctions and the first junction is supported by 2 species while the second one is supported by 3, the conservation score is 3.

4.5.7 Mapping of PSFs to Structure

A PSI-BLAST [104] position specific matrix (PSSM) is complied for the protein sequence of each major isoform, by searching against the Uniprot [105] database for 3 rounds. The PSSMs were then used to search the RCSB [106] protein sequence database for potential homologous templates using PSI-BLAST with an E-score cutoff of 10e-5. The location of each exon in the 3D structure is obtained by mapping the protein segment corresponding to the exon onto the alignment. The PSF coverage score was calculated as the fraction of all residues in the PSF that are covered by a structural template. Only PSFs that are 95% covered by a structural template were used in this study

4.5.8 Calculation of Structural Properties

We deleted the atomic coordinates of protein fragments corresponding to Deletion PSFs from the structural templates and calculated various statistics.CCP4 [107] was used to calculate the exposed surface area of all atoms in the original templates and for all atoms in the modified templates formed by deletion of PSFs. The newly exposed area is calculated as the sum over all atoms that were previously buried but now exposed (buried defined as zero surface area). The newly exposed hydrophobic area is the sum of all contribution from carbon atoms to newly exposed area. The number of lost contacts is the

number of contacts in an original template that no longer exist in the modified template (contact defined as distance between atoms of less than $6\mathring{A}$). Deletion end-to-end distance was calculated as the distance between C-alpha atoms of residues at each end of a deletion.

Chapter 5

Chapter 5. Conclusion

Based on the frequency of alternative splicing events and their consequence on protein sequences and structures, we conclude that mistakes in the selection of splice sites largely determins the number of observed isoforms for a given gene. As was shown in the last chapter, isoforms resulting from mistakes in the selection of splice sites are unlikely to result in a production of stable proteins, and therefore unlikely to result in functional protein products. We do not know exactly what fraction of all isoforms are functional, but our expectation is that this number is not much higher than fraction of isoforms conserved across multiple species.

Although, we have not considered the long term consequence of alternative splicing on the evolution of new functionality, the process by which alternative splicing transcripts come to code for new functionality is of great interest. It has been argued that alternative splicing can create opportunities for evolution of new functionality [108, 30, 32]. For example, Letunic et al. have found that approximately 10% of all genes in H.sapiens, D.melanogaster and C.elegans contain exon duplications and that these exons are under increased selection pressure (reduced Ka/Ks ratio), suggesting that mutualy exclusive selection of duplicated exons provides a framework for evolution of a novel functionality [108].

More recently analysis of mutation rates in mammalian genes by Xing et al has shown that Ka/Ks ratios are correlated to the expression level of the exons [32]. They found that alternatively spliced exons expressed at low abundance levels are under re-

duced selection pressure (higher Ka/Ks) compared to constitutive exons and highly expressed alternative exons. Remarkably, the reduction in selection pressure with decreased expression level was not due to a significant increase in amino acid substation rate Ka, but due to a large decrease in synonymous substitution Ks. Reduced Ks is seen as evidence for an increased selection pressure on a regulatory regions within RNA. Combined with observation that there is an increased preservation of coding frames by alternative splicing, the authors argue that these observations are consistent with the hypothesis that alternative splicing provides a mechanism for evolution of new functionality by creating evolutionary ”hot-spots - localized regions of a gene structure with an accelerated rate of evolution.

Intron regions around alternativly spliced exons have also been shown to be under increased selection pressure (reduced nucleotide substitution rates) by a number researchers [109, 110, 111]. Sorek et al. compared intronic sequences of human and mouse genomes flanking alternative exons and found that these regions tend to have a higher conservation level compared to constitutive exons. Remarkably they also find that conserved regions around alternative exons tend to be longer than constitutive exons. Thus, it is argued that there is increased selection pressure on introns around alternative exons due to the increased need for regulatory control of these exons.

A general theory that has emerging from these observations is that alternative splicing provides a neutral pathway for exploring the functionality landscape (reviewed [112]). The validity of this hypothesis is hard to judge without knowing the costs of making mistakes. We show in Chapter 4 that one of the costs associated with increased alternative splicing is increased production of unstable protein folds, which if not degraded, will have toxic consequences on a cell. Based on analysis of frequency of alternative splicing, we also conclude that splicing error rates are tuned for each gene, to minimize production of non viable protein products to such a level that they do not interfere with the normal

function of a gene. Once functional and toxic constrains on splicing are satisfied, there will be little selection pressure to further reduce frequency of mistakes, allowing for a production of rich diversity of the trial transcripts.

Chapter 6

Common Methods Appendix

6.1 Data Sources

The human genome sequence[113] was downloaded from NCBI (NCBI Human Genome Build 35). The transcript data was obtained from Refseq [70] (Release 17; May 2006; 29,475 sequences), Unigene [70](May 2006; 6,586 ,000 sequences), and H-InvDB [58] (Release 3.0;4 49,186 sequences). The location of genes to chromosomes was taken from Refseq database annotation. Information about homologous genes in other species was obtained from the NCBI Homologene Database [70] (Release 48, May 2006).

For each gene, all sequences were aligned to a human genomic contig using the sim4 algorithm [114] and then checked for alignment errors (see list of rules below).

6.2 Alignment Quality Control

The following five rules are used to identify sequences containing alignment and sequencing errors.

1. All implied splice sites must conform to the spliceosome pattern -'GT/AG'.
2. All exons must have greater than 90% identity with the corresponding genomic sequence.
3. Alignment to genomic sequence must not contain any missing segments (see Figure 6.1).
4. The sequence around exon junctions (6 nucleotides into each exon) must have 100% identity with the corresponding contig.

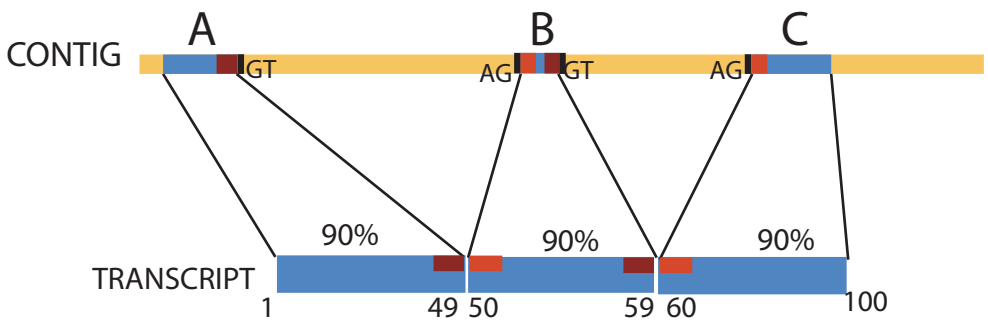

Figure 6.1: Alignment Quality Controls. A hypothetical 100 nucleotide transcript with three exons (A, B, C) is aligned to a corresponding genomic sequence. Introns are verified to conform to the 'GT-AG' motif. All introns must be more than 30 nucleotides in length. All exons must align with at least 90% identity to the contig. The alignment must not contain any missing segments. For example, exon A stops the alignment at position 49, and so exon B must start the alignment at position 50. A sequence on each side of an exon-exon junction (6 nucleotides on each side - highlighted in red) must align to the contig with 100% identity and no gaps). If the application calls for a full length alignment, the 5' end of the transcript must start alignment at the first nucleotide, and the 3' end of the transcript must end alignment at the last nucleotide (position 100 in this example).

5. The cDNA must not contain any introns of size less than 30 nucleotides

Two additional filters were applied to minor isoforms: 1. Minor isoforms must share at least one exon with the corresponding major isoform (overlap of greater than 1 nucleotide). 2. Minor isoforms must not contain an intron retention event relative to the major isoform.

6.3 Selection of Major Isoform

For each gene, we identified one of the cDNAs as the major isoform - that is, the isoform whose splicing patterns are most frequently observed across all Unigene EST

libraries. The exon structure of major isoforms is used as a reference to which the exon structures of minor isoforms are compared. To determine major isoforms, sequences are sorted using the following procedure: First we created a list of introns and all sequences that are associated with those introns. For each intron in a cDNA, we calculate the number of EST sequences and number of unique EST libraries that contain this intron. For each cDNA we then compute three values: sequence length, number of ESTs containing one or more of its introns, and the number of unique EST libraries containing any of its introns. Finally, we sort the cDNAs using these values in the following order: 1. Number of unique EST libraries. 2. Total number of ESTs. 3. Sequence Length. The top ranking sequence is selected as the major isoform.

Bibliography

[1] Custdio, N., Carmo-Fonseca, M., Geraghty, F., Pereira, H. S., Grosveld, F., and Antoniou, M. (May, 1999) Inefficient processing impairs release of RNA from the site of transcription.. *EMBO J,* **18**(10), 2855–2866.

[2] Carter, M. G., Sharov, A. A., VanBuren, V., Dudekula, D. B., Carmack, C. E., Nelson, C., and Ko, M. S. (2005) Transcript copy number estimation using a mouse whole-genome oligonucleotide microarray. *Genome Biol,* **6**(7), R61.

[3] Kornblihtt, A. R., de laMata, M., Fededa, J. P., Munoz, M. J., and Nogues, G. (Oct, 2004) Multiple links between transcription and splicing.. *RNA,* **10**(10), 1489–1498.

[4] Jurica, M. S. and Moore, M. J. (Jul, 2003) Pre-mRNA splicing: awash in a sea of proteins.. *Mol Cell,* **12**(1), 5–14.

[5] Guig, R., Flicek, P., Abril, J. F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V. B., Birney, E., Castelo, R., Eyras, E., Ucla, C., Gingeras, T. R., Harrow, J., Hubbard, T., Lewis, S. E., and Reese, M. G. (2006) EGASP: the human ENCODE Genome Annotation Assessment Project.. *Genome Biol,* **7 Suppl 1**, S2.1–S231.

[6] Black, D. L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem,* **72**, 291–336.

[7] Sanford, J. R., Ellis, J., and Cáceres, J. F. (Jun, 2005) Multiple roles of arginine/serine-rich splicing factors in RNA processing.. *Biochem Soc Trans,* **33**(Pt 3), 443–446.

[8] Krecic, A. M. and Swanson, M. S. (Jun, 1999) hnRNP complexes: composition, structure, and function.. *Curr Opin Cell Biol,* **11**(3), 363–371.

[9] Smith, C. W. and Valcrcel, J. (Aug, 2000) Alternative pre-mRNA splicing: the logic of combinatorial control.. *Trends Biochem Sci,* **25**(8), 381–388.

[10] Takeda, J.-I., Suzuki, Y., Nakao, M., Barrero, R., Koyanagi, K., Jin, L., Motono, C., Hata, H., Isogai, T., Nagai, K., Otsuki, T., Kuryshev, V., Shionyu, M., Yura, K., Go, M., Thierry-Mieg, J., Thierry-Mieg, D., Wiemann, S., Nomura, N., Sugano, S., Gojobori, T., and Imanishi (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs.. *Nucleic acids research.,* **34**(14), 3917–28.

[11] Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E., and Zipursky, S. L. (Jun, 2000) Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity.. *Cell,* **101**(6), 671–684.

[12] Celotto, A. M. and Graveley, B. R. (2001) Alternative splicing of the Drosophila Dscam pre-mRNA is both temporally and spatially regulated. *Genetics,* **159**(2), 599–608.

[13] Missler, M. and Sdhof, T. C. (Jan, 1998) Neurexins: three genes and 1001 products.. *Trends Genet,* **14**(1), 20–26.

[14] Zhu, J., Shendure, J., Mitra, R. D., and Church, G. M. (2003) Single molecule profiling of alternative pre-mRNA splicing. *Science,* **301**(5634), 836–8.

[15] Penalva, L. O. and Sanchez, L. (2003) RNA binding protein sex-lethal (Sxl) and control of Drosophila sex determination and dosage compensation. *Microbiol Mol Biol Rev,* **67**(3), 343–59, table of contents.

[16] Jensen, K. B., Dredge, B. K., Stefani, G., Zhong, R., Buckanovich, R. J., Okano, H. J., Yang, Y. Y., and Darnell, R. B. (Feb, 2000) Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability.. *Neuron,* **25**(2), 359–371.

[17] Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.-S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., Zeeberg, B. R., Kane, D., Weinstein, J. N., Blume, J., and Darnell, R. B. (Aug, 2005) Nova regulates brain-specific splicing to shape the synapse.. *Nat Genet,* **37**(8), 844–852.

[18] Mironov, A. A., Fickett, J. W., and Gelfand, M. S. (1999) Frequent Alternative Splicing of Human Genes. *Genome Res.,* **9**(12), 1288–1293.

[19] Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. (2000) EST comparison indicates 38splice forms. *FEBS Letters,* **474**(1), 83–86.

[20] Modrek, B., Resch, A., Grasso, C., and Lee, C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res,* **29**(13), 2850–9.

[21] Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., and Shoemaker, D. D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science,* **302**(5653), 2141–4.

[22] Black, D. L. (Oct, 2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology.. *Cell,* **103**(3), 367–370.

[23] Graveley, B. R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet,* **17**(2), 100–7.

[24] Maniatis, T. and Tasic, B. (Jul, 2002) Alternative pre-mRNA splicing and proteome expansion in metazoans.. *Nature,* **418**(6894), 236–243.

[25] Modrek, B. and Lee, C. J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet,* **34**(2), 177–80.

[26] Kan, Z., Garrett-Engele, P. W., Johnson, J. M., and Castle, J. C. (2005) Evolutionarily conserved and diverged alternative splicing events show different expression and functional profiles. *Nucleic Acids Research,* **33**(17), 5659.

[27] Nurtdinov, R. N., Artamonova, I., Mironov, A. A., and Gelfand, M. S. (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum Mol Genet,* **12**(11), 1313–20.

[28] Xu, Q., Modrek, B., and Lee (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome.. *Nucleic acids research.,* **30**(17), 3754–66.

[29] Yeo, G., Holste, D., Kreiman, G., and Burge, C. B. (2004) Variation in alternative splicing across human tissues.. *Genome Biol,* **5**(10), R74.

[30] Boue, S., Letunic, I., and Bork, P. (2003) Alternative splicing and evolution. *Bioessays,* **25**(11), 1031–4.

[31] Resch, A., Xing, Y., Alekseyenko, A., Modrek, B., and Lee (2004) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation.. *Nucleic acids research.,* **32**(4), 1261–9.

[32] Xing, Y. and Lee (2005) Colloquium Paper: Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proceedings of the National Academy of Sciences,* **102**(38), 13526.

[33] Lewis, B., Green, R., and Brenner, S. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans.. *Proceedings of the National Academy of Sciences of the United States of America.,* **100**(1), 189–92.

[34] Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A. L., Mohammad, N., Babak, T., Siu, H., Hughes, T. R., Morris, Q. D., Frey, B. J., and Blencowe, B. J. (Dec, 2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform.. *Mol Cell,* **16**(6), 929–941.

[35] Pan, Q., Saltzman, A. L., Kim, Y. K., Misquitta, C., Shai, O., Maquat, L. E., Frey, B. J., and Blencowe, B. J. (Jan, 2006) Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression.. *Genes Dev,* **20**(2), 153–158.

[36] Hillman, R. T., Green, R. E., and Brenner, S. E. (2004) An unappreciated role for RNA surveillance.. *Genome Biol,* **5**(2), R8.

[37] Matlin, A. J., Clark, F., and Smith, C. W. J. (May, 2005) Understanding alternative splicing: towards a cellular code.. *Nat Rev Mol Cell Biol,* **6**(5), 386–398.

[38] Kan, Z., States, D., and Gish, W. (2002) Selecting for functional alternative splices in ESTs. *Genome Res,* **12**(12), 1837–45.

[39] Sorek, R., Shamir, R., and Ast (2004) How prevalent is functional alternative splicing in the human genome?. *Trends in genetics : TIG.,* **20**(2), 68–71.

[40] Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D. A., Hayashizaki, Y., Gaasterland, T., Group, R. I. K. E. N. G., and Members, G. S. L. (Jun, 2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome.. *Genome Res,* **13**(6B), 1290–1300.

[41] Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., deBono, B., Gatta, G. D., diBernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P. T., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Babu, M. M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schnbach, C., Sekiguchi, K., Semple, C. A. M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., vanNimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M.,

Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., Consortium, F. A. N. T. O. M., Group, R. I. K. E. N. G. E. R., and Group), G. S. G. G. N. P. C. (Sep, 2005) The transcriptional landscape of the mammalian genome.. *Science,* **309**(5740), 1559–1563.

[42] Landry, J.-R., Mager, D. L., and Wilhelm, B. T. (Nov, 2003) Complex controls: the role of alternative promoters in mammalian genomes.. *Trends Genet,* **19**(11), 640–648.

[43] Tian, B., Hu, J., Zhang, H., and Lutz, C. S. (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes.. *Nucleic Acids Res,* **33**(1), 201–212.

[44] Ladomery, M. (Oct, 1997) Multifunctional proteins suggest connections between transcriptional and post-transcriptional processes.. *Bioessays,* **19**(10), 903–909.

[45] Neugebauer, K. M. (Oct, 2002) On the importance of being co-transcriptional.. *J Cell Sci,* **115**(Pt 20), 3865–3871.

[46] Maniatis, T. and Reed, R. (2002) An extensive network of coupling among gene expression machines. *Nature,* **416**(6880), 499–506.

[47] Pagani, F., Stuani, C., Zuccato, E., Kornblihtt, A. R., and Baralle, F. E. (2003) Promoter Architecture Modulates CFTR Exon 9 Skipping. *J. Biol. Chem.,* **278**(3), 1511–1517.

[48] Nogues, G., Munoz, M. J., and Kornblihtt, A. R. (2003) Influence of Polymerase II Processivity on Alternative Splicing Depends on Splice Site Strength. *J. Biol. Chem.,* **278**(52), 52166–52171.

[49] Vagner, S., Vagner, C., and Mattaj, I. W. (2000) The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3'-end processing and splicing. *Genes Dev.,* **14**(4), 403–413.

[50] Castelo-Branco, P., Furger, A., Wollerton, M., Smith, C., Moreira, A., and Proudfoot, N. (2004) Polypyrimidine Tract Binding Protein Modulates Efficiency of Polyadenylation. *Mol. Cell. Biol.,* **24**(10), 4174–4183.

[51] Furger, A., Binnie, Justin M. O'Sullivan, A., Lee, B. A., and Proudfoot, N. J. (2002) Promoter proximal splice sites enhance transcription. *Genes Dev.,* **16**(21), 2792–2799.

[52] Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engstrm, P. G., Frith, M. C., Forrest, A. R. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A.,

and Hayashizaki, Y. (Jun, 2006) Genome-wide analysis of mammalian promoter architecture and evolution.. *Nat Genet,* **38**(6), 626–635.

[53] Frilander, M. J. and Steitz, J. A. (Apr, 1999) Initial recognition of U12-dependent introns requires both U11/5' splice-site and U12/branchpoint interactions.. *Genes Dev,* **13**(7), 851–863.

[54] Alioto, T. S. (Jan, 2007) U12DB: a database of orthologous U12-type spliceosomal introns.. *Nucleic Acids Res,* **35**(Database issue), D110–D115.

[55] Schneider, T. D. and Stephens, R. M. (Oct, 1990) Sequence logos: a new way to display consensus sequences.. *Nucleic Acids Res,* **18**(20), 6097–6100.

[56] Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (Jun, 2004) WebLogo: a sequence logo generator.. *Genome Res,* **14**(6), 1188–1190.

[57] Robberson, B. L., Cote, G. J., and Berget, S. M. (Jan, 1990) Exon definition may facilitate splice site selection in RNAs with multiple exons.. *Mol Cell Biol,* **10**(1), 84–94.

[58] Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. O., Barrero, R. A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M., Yura, K., Miyazaki, S., Ikeo, K., Homma, K., Kasprzyk, A., Nishikawa, T., Hirakawa, M., Thierry-Mieg, J., Thierry-Mieg, D., Ashurst, J., Jia, L., Nakao, M., Thomas, M. A., Mulder, N., Karavidopoulou, Y., Jin, L., Kim, S., Yasuda, T., Lenhard, B., Eveno, E., Suzuki, Y., Yamasaki, C., ichiTakeda, J., Gough, C., Hilton, P., Fujii, Y., Sakai, H., Tanaka, S., Amid, C., Bellgard, M., deFatima Bonaldo, M., Bono, H., Bromberg, S. K., Brookes, A. J., Bruford, E., Carninci, P., Chelala, C., Couillault, C., deSouza, S. J., Debily, M.-A., Devignes, M.-D., Dubchak, I., Endo, T., Estreicher, A., Eyras, E., Fukami-Kobayashi, K., Gopinath, G. R., Graudens, E., Hahn, Y., Han, M., Han, Z.-G., Hanada, K., Hanaoka, H., Harada, E., Hashimoto, K., Hinz, U., Hirai, M., Hishiki, T., Hopkinson, I., Imbeaud, S., Inoko, H., Kanapin, A., Kaneko, Y., Kasukawa, T., Kelso, J., Kersey, P., Kikuno, R., Kimura, K., Korn, B., Kuryshev, V., Makalowska, I., Makino, T., Mano, S., Mariage-Samson, R., Mashima, J., Matsuda, H., Mewes, H.-W., Minoshima, S., Nagai, K., Nagasaki, H., Nagata, N., Nigam, R., Ogasawara, O., Ohara, O., Ohtsubo, M., Okada, N., Okido, T., Oota, S., Ota, M., Ota, T., Otsuki, T., Piatier-Tonneau, D., Poustka, A., Ren, S.-X., Saitou, N., Sakai, K., Sakamoto, S., Sakate, R., Schupp, I., Servant, F., Sherry, S., Shiba, R., Shimizu, N., Shimoyama, M., Simpson, A. J., Soares, B., Steward, C., Suwa, M., Suzuki, M., Takahashi, A., Tamiya, G., Tanaka, H., Taylor, T., Terwilliger, J. D., Unneberg, P., Veeramachaneni, V., Watanabe, S., Wilming, L., Yasuda, N., Yoo, H.-S., Stodolsky, M., Makalowski, W., Go, M., Nakai, K., Takagi, T., Kanehisa, M., Sakaki, Y., Quackenbush, J., Okazaki, Y., Hayashizaki, Y., Hide, W., Chakraborty, R., Nishikawa, K., Sugawara, H., Tateno, Y., Chen, Z., Oishi, M., Tonellato, P., Apweiler, R., Okubo, K., Wagner, L., Wiemann, S., Strausberg, R. L., Isogai, T., Auffray, C., Nomura, N., Gojobori, T., and Sugano, S.

(Jun, 2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones.. *PLoS Biol,* **2**(6), e162.

[59] Modrek, B. and Lee, C. (2002) A genomic view of alternative splicing. *Nature Genetics,* **30**(1), 13.

[60] Kriventseva, E. V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M. S., and Sunyaev, S. (2003) Increase of functional diversity by alternative splicing. *Trends Genet,* **19**(3), 124–8.

[61] Magen, A. and Ast, G. (2005) The importance of being divisible by three in alternative splicing. *Nucleic Acids Research,* **33**(17), 5574.

[62] Thanaraj, T. A., Clark, F., and Muilu, J. (2003) Conservation of human alternative splice events in mouse. *Nucleic Acids Res,* **31**(10), 2544–52.

[63] Sorek, R., Dror, G., and Shamir, R. (2006) Assessing the number of ancestral alternatively spliced exons in the human genome.. *BMC Genomics,* **7**, 273.

[64] Pan, Q., Bakowski, M. A., Morris, Q., Zhang, W., Frey, B. J., Hughes, T. R., and Blencowe, B. J. (Feb, 2005) Alternative splicing of conserved exons is frequently species-specific in human and mouse.. *Trends Genet,* **21**(2), 73–77.

[65] Kondrashov, F. A. and Koonin, E. V. (2003) Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet,* **19**(3), 115–9.

[66] Rehwinkel, J., Letunic, I., Raes, J., Bork, P., and Izaurralde, E. (Oct, 2005) Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets.. *RNA,* **11**(10), 1530–1544.

[67] Gilbert, W. (Feb, 1978) Why genes in pieces?. *Nature,* **271**(5645), 501.

[68] Ermakova, E. O., Nurtdinov, R. N., and Gelfand, M. S. (2006) Fast rate of evolution in alternatively spliced coding regions of mammalian genes.. *BMC Genomics,* **7**, 84.

[69] Bonaldo, M. F., Lennon, G., and Soares, M. B. (Sep, 1996) Normalization and subtraction: two approaches to facilitate gene discovery.. *Genome Res,* **6**(9), 791–806.

[70] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Helmberg, W., Kapustin, Y., Kenton, D. L., Khovayko, O., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Pruitt, K. D., Schuler, G. D., Schriml, L. M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Suzek, T. O., Tatusov, R., Tatusova, T. A., Wagner, L., and Yaschenko, E. (Jan, 2006) Database resources of the National Center for Biotechnology Information.. *Nucleic Acids Res,* **34**(Database issue), D173–D180.

[71] Strausberg, R. L. (Sep, 2001) The Cancer Genome Anatomy Project: new resources for reading the molecular signatures of cancer.. *J Pathol,* **195**(1), 31–40.

[72] Permutt, A. UNIGENE Library 8840- large single tissue EST library derived from purified pancreatic islet cells..

[73] Goldberg, A. L. (2003) Protein degradation and protection against misfolded or damaged proteins. *Nature,* **426**(6968), 895–9.

[74] Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. (Oct, 2005) Why highly expressed proteins evolve slowly.. *Proc Natl Acad Sci U S A,* **102**(40), 14338–14343.

[75] Ellis, R. J. and Pinheiro, T. J. T. (Apr, 2002) Medicine: danger–misfolding proteins.. *Nature,* **416**(6880), 483–484.

[76] Neverov, A. D., Artamonova, I., Nurtdinov, R. N., Frishman, D., Gelfand, M. S., and Mironov, A. A. (2005) Alternative splicing and protein function. *BMC Bioinformatics,* **6**, 266.

[77] Drummond, D. A., Raval, A., and Wilke, C. O. (Feb, 2006) A single determinant dominates the rate of yeast protein evolution.. *Mol Biol Evol,* **23**(2), 327–337.

[78] Yoon, S. S., Segal, N. H., Park, P. J., Detwiller, K. Y., Fernando, N. T., Ryeom, S. W., Brennan, M. F., and Singer, S. (Oct, 2006) Angiogenic profile of soft tissue sarcomas based on analysis of circulating factors and microarray gene expression.. *J Surg Res,* **135**(2), 282–290.

[79] Pertea, M., Lin, X., and Salzberg, S. L. (Mar, 2001) GeneSplicer: a new computational method for splice site prediction.. *Nucleic Acids Res,* **29**(5), 1185–1190.

[80] Fairbrother, W. G., Yeh, R. F., Sharp, P. A., and Burge, C. B. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science,* **297**(5583), 1007–13.

[81] Akaike, H. (1981) Likelihood of a model and information criteria. *Journal of Econometrics,* **16**(1), 3–14.

[82] Graveley, B. R. (Feb, 2001) Alternative splicing: increasing diversity in the proteomic world.. *Trends Genet,* **17**(2), 100–107.

[83] Chen, B. E., Kondo, M., Garnier, A., Watson, F. L., Pettmann-Holgado, R., Lamar, D. R., and Schmucker, D. (May, 2006) The molecular diversity of Dscam is functionally required for neuronal wiring specificity in Drosophila.. *Cell,* **125**(3), 607–620.

[84] Romero, P. R., Zaidi, S., Fang, Y. Y., Uversky, V. N., Radivojac, P., Oldfield, C. J., Cortese, M. S., Sickmeier, M., LeGall, T., Obradovic, Z., and Dunker, A. K. (May,

2006) Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms.. *Proc Natl Acad Sci U S A,* **103**(22), 8390–8395.

[85] Veldhoen, N., Metcalfe, S., and Milner, J. (Nov, 1999) A novel exon within the mdm2 gene modulates translation initiation in vitro and disrupts the p53-binding domain of mdm2 protein.. *Oncogene,* **18**(50), 7026–7033.

[86] Lamba, J. K., Adachi, M., Sun, D., Tammur, J., Schuetz, E. G., Allikmets, R., and Schuetz, J. D. (Jan, 2003) Nonsense mediated decay downregulates conserved alternatively spliced ABCC4 transcripts bearing nonsense codons.. *Hum Mol Genet,* **12**(2), 99–109.

[87] Winter, J., Lehmann, T., Krauss, S., Trockenbacher, A., Kijas, Z., Foerster, J., Suckow, V., Yaspo, M.-L., Kulozik, A., Kalscheuer, V., Schneider, R., and Schweiger, S. (May, 2004) Regulation of the MID1 protein function is fine-tuned by a complex pattern of alternative splicing.. *Hum Genet,* **114**(6), 541–552.

[88] Wittmann, J., Hol, E. M., and Jck, H.-M. (Feb, 2006) hUPF2 silencing identifies physiologic substrates of mammalian nonsense-mediated mRNA decay.. *Mol Cell Biol,* **26**(4), 1272–1287.

[89] Blencowe, B. J. (Jul, 2006) Alternative splicing: new insights from global analyses.. *Cell,* **126**(1), 37–47.

[90] Yeo, G. W., Nostrand, E. V., Holste, D., Poggio, T., and Burge, C. B. (Feb, 2005) Identification and analysis of alternative splicing events conserved in human and mouse.. *Proc Natl Acad Sci U S A,* **102**(8), 2850–2855.

[91] Pruitt, K., Tatusova, T., and Maglott, D. (2004) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research,* **33**(database issue), 501.

[92] Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Tatusova, T. A., and Wagner, L. (Jan, 2003) Database resources of the National Center for Biotechnology.. *Nucleic Acids Res,* **31**(1), 28–33.

[93] Nagy, E. and Maquat, L. E. (Jun, 1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance.. *Trends Biochem Sci,* **23**(6), 198–199.

[94] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (Jan, 2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.. *Nucleic Acids Res,* **33**(Database issue), D514–D517.

[95] Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., Abeysinghe, S., Krawczak, M., and Cooper, D. N. (Jun, 2003) Human Gene Mutation Database (HGMD): 2003 update.. *Hum Mutat,* **21**(6), 577–581.

[96] Pagon, R. A., Tarczy-Hornoch, P., Baskin, P. K., Edwards, J. E., Covington, M. L., Espeseth, M., Beahler, C., Bird, T. D., Popovich, B., Nesbitt, C., Dolan, C., Marymee, K., Hanson, N. B., Neufeld-Kaiser, W., Grohs, G. M., Kicklighter, T., Abair, C., Malmin, A., Barclay, M., and Palepu, R. D. (May, 2002) GeneTests-GeneClinics: genetic testing information for a growing audience.. *Hum Mutat,* **19**(5), 501–509.

[97] ichiTakeda, J., Suzuki, Y., Nakao, M., Barrero, R. A., Koyanagi, K. O., Jin, L., Motono, C., Hata, H., Isogai, T., Nagai, K., Otsuki, T., Kuryshev, V., Shionyu, M., Yura, K., Go, M., Thierry-Mieg, J., Thierry-Mieg, D., Wiemann, S., Nomura, N., Sugano, S., Gojobori, T., and Imanishi, T. (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs.. *Nucleic Acids Res,* **34**(14), 3917–3928.

[98] Liu, S. and Altman, R. B. (Aug, 2003) Large scale study of protein domain distribution in the context of alternative splicing.. *Nucleic Acids Res,* **31**(16), 4828–4835.

[99] Resch, A., Xing, Y., Modrek, B., Gorlick, M., Riley, R., and Lee, C. (2004) Assessing the impact of alternative splicing on domain interactions in the human proteome.. *J Proteome Res,* **3**(1), 76–83.

[100] Homma, K., Kikuno, R. F., Nagase, T., Ohara, O., and Nishikawa, K. (Nov, 2004) Alternative splice variants encoding unstable protein domains exist in the human brain.. *J Mol Biol,* **343**(5), 1207–1220.

[101] Wang, P., Yan, B., Guo, J.-T., Hicks, C., and Xu, Y. (Dec, 2005) Structural genomics analysis of alternative splicing and application to isoform structure modeling.. *Proc Natl Acad Sci U S A,* **102**(52), 18920–18925.

[102] Kozak, M. (1992) Regulation of translation in eukaryotic systems.. *Annu Rev Cell Biol,* **8**, 197–225.

[103] Wang, X.-Q. and Rothnagel, J. A. (2004) 5'-untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation.. *Nucleic Acids Res,* **32**(4), 1382–1391.

[104] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* **25**(17), 3389–402.

[105] Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J.,

Mazumder, R., O'Donovan, C., Redaschi, N., and Suzek, B. (Jan, 2006) The Universal Protein Resource (UniProt): an expanding universe of protein information.. *Nucleic Acids Res,* **34**(Database issue), D187–D191.

[106] Deshpande, N., Addess, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z., Green, R. K., Flippen-Anderson, J. L., Westbrook, J., Berman, H. M., and Bourne, P. E. (Jan, 2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema.. *Nucleic Acids Res,* **33**(Database issue), D233–D237.

[107] Winn, M. D., Ashton, A. W., Briggs, P. J., Ballard, C. C., and Patel, P. (Nov, 2002) Ongoing developments in CCP4 for high-throughput structure determination.. *Acta Crystallogr D Biol Crystallogr,* **58**(Pt 11), 1929–1936.

[108] Letunic, I., Copley, R. R., and Bork, P. (Jun, 2002) Common exon duplication in animals and its role in alternative splicing.. *Hum Mol Genet,* **11**(13), 1561–1567.

[109] Sorek, R. and Ast, G. (Jul, 2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse.. *Genome Res,* **13**(7), 1631–1637.

[110] Sugnet, C. W., Srinivasan, K., Clark, T. A., O'Brien, G., Cline, M. S., Wang, H., Williams, A., Kulp, D., Blume, J. E., Haussler, D., and Ares, M. (Jan, 2006) Unusual intron conservation near tissue-regulated exons found by splicing microarrays.. *PLoS Comput Biol,* **2**(1), e4.

[111] Xing, Y., Wang, Q., and Lee, C. (Jul, 2006) Evolutionary divergence of exon flanks: a dissection of mutability and selection.. *Genetics,* **173**(3), 1787–1791.

[112] Xing, Y. and Lee, C. (Jul, 2006) Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes.. *Nat Rev Genet,* **7**(7), 499–509.

[113] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P.,

Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., and Frazier, M., e. a. (2001) Initial sequencing and analysis of the human genome. *Nature,* **409**(6822), 860–921.

[114] Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res,* **8**(9), 967–74.