The Evolution and Specificity of RNA Splicing

by

Brad Aaron Friedman

B.S., University of Illinois at Urbana-Champaign, 1999
B.M., University of Illinois at Urbana-Champaign, 2000
C.A.S.M., University of Cambridge, 2001

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2006

Author .....
Department of Mathematics
August 11, 2006

Certified by .....
Christopher B. Burge
Associate Professor of Biology
Department of Biology
Thesis Supervisor

Accepted by .....
Alar Toomre
Chairman, Applied Mathematics Committee

Accepted by .....
Pavel I. Etingof
Chairman, Department Committee on Graduate Students

The Evolution and Specificity of RNA Splicing

by

Brad Aaron Friedman

Submitted to the Department of Mathematics
on August 11, 2006, in partial fulfillment of the
requirements for the degree of
DOCTOR OF PHILOSOPHY

Abstract

The majority of human genes are not encoded in contiguous segments in the genome but are rather punctuated by long interruptions known as *introns*. These introns are copied from generation to generation, and even from cell to cell as a person grows from an embryo into an adult. Each time a gene is activated, the cell must first accurately excise all the introns in a process known as *splicing*.

This excision is determined by the sequence of the gene, but in a complicated way that is not fully understood. By analyzing gene sequences we can learn about how cells decide which sequences to splice. We have developed two new mathematical models, one for the end of introns, and another for long distance interactions between different parts of genes, that expose previously unknown elements potentially involved in the splicing reaction.

However their boundaries are determined, introns are very ancient: although they are absent from bacteria they are found in almost all protists, fungi, plants and animals. It is therefore of great interest to explain their evolutionary origins. We have developed a probabilistic model for the evolution of introns and used it to perform a genome-wide analysis of the patterns of intron conservation in four euascomycete fungi, establishing that intron gain and loss are constantly reshaping the distribution of introns in genes.

Thesis Supervisor: Christopher B. Burge
Title: Associate Professor of Biology, Department of Biology

Acknowledgments

I would like to first thank my advisor Christopher Burge for giving me the chance to study in a field with which I was entirely unfamiliar. Without his support my research would have been impossible. I hope that some of his scientific philosophy has been transmitted to me through our many discussions over the past four years.

Every member of the Burge lab has been, at some point or another, a teacher or friend to me. I feel a great debt to my colleagues. I would like to thank Will Fairbrother, Luba Katz and Mehdi Yahyanejad for welcoming me to the group during my first years, and in particular Mehdi for many interesting conversations on a variety of scientific and other topics. I thank Uwe Ohler for tips on the technical issues of bioinformatics as well as career advice. Cydney Nielsen deserves my thanks for laying the foundations for the first project I completed in the lab. In my last years in the lab, Michael Stadler, Grace Xiao and Rickard Sandberg have been constant companions and invaluable resources. I cannot thank Zefeng Wang and Noam Shomron enough. If I have learned anything useful about experimental molecular biology it has been from them.

I thank Jan Vondrak, Fumei Lam, Ashraf Ibrahim and Daniel Quest, friends from other departments and universities, for answering questions and listening to ideas of mine, some too technical to capture the attention of biologists.

I thank Bonnie Berger who has served as my internal math department advisor, facilitating my transition to from pure mathematics to computational biology. I also thank James Galagan of the Broad Institute. He spent many hours perfecting our manuscript for publication, and since then has always made time to advise me on scientific and career issues.

I am indebted to our present and past lab administrators, Brenda Pepe and Alison Hearn, and to Linda Okun, the math department graduate student administrator. These three allowed me to focus on intellectual pursuits, making my graduate career smooth in all other respects.

I thank my parents, Arthur and Erica Friedman, and my sisters, Monica and

Stacy Friedman, for their support and encouragement. Finally I thank my wife, Irene Friedman, and our son Noah. They have given science and every breath I take a new meaning.

Brad Friedman

Cambridge, Massachusetts

August 10, 2006

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> The role of enzymes in the life of the cell is to mediate the transformation of the simple chemical world of the environment into living protoplasm. In effect, they guide molecules across the threshold of life.[13]

Biochemistry was born when scientists in the nineteenth century studying fermentation discovered that the conversion of sugar to ethanol was catalyzed by enzymes, molecular machines formed by living yeast cells. Such machines are responsible for every cellular process in the human body, from processing the air we breathe to helping us digest our food to enabling us to have thoughts. We now know that the vast majority of these machines are proteins, and that each one arises from its own gene.

The program of gene expression is therefore tightly regulated in every cell. Not only does each protein need to be synthesized correctly in order to function, but it needs to be synthesized in the right cells under the right conditions. Proteins are linear chains of repetitive subunits, each subunit being one of twenty amino acids, and it is this sequence of amino acids that determines all the chemical properties of the protein—for example, which reaction it will catalyze, how quickly it will degrade, into what shape it will fold, and whether or not it will aggregate. Errors in the sequence, sometimes even a single incorrect amino acid, can cause the protein to malfunction and lead to serious diseases (anemia, cystic fibrosis, and Alzheimer's disease to name a few). On the other hand, even proteins synthesized correctly can be deadly when

Figure 1-1: A schematic example of four bases of double-stranded DNA. The backbone is shown in gray and the bases in black.

they are made in the wrong cells or at the wrong time. For example, mis-regulation of many cell-cycle proteins causes cancer.

The sequence of every protein is encoded by a gene, and these genes are inherited from generation to generation. Aside from the protein sequence, each gene contains signals for the cell to use that control its expression. Understanding *gene expression*, the pathway that starts with genes and ends with correctly regulated proteins, is therefore of vital importance to both biology and medicine.

1.1 The Life Cycle of a Gene

Genes are encoded in DNA in a four letter chemical alphabet (adenosine, guanosine, cytidine and thymine, abbreviated `A`, `C`, `G` and `T` respectively) in long linear molecules called chromosomes. These chromosomes are two-stranded polymers made of these subunits, where the residues from each polymer are paired up like half-ties of a railroad track, with the bases `C` and `G` always paired, and `A` and `T` always paired (Figure 1-1). The strands each have an orientation beginning from the so-called "5′ end" and ending at the "3′ end". In humans, 24 distinct chromosomes make up the entire genome, 22 autosomes and 2 sex chromosomes, and they range in length from about 50 million to 250 million base pairs, totaling about 3 billion bases.

Although the sequence of one strand can be determined from the other by reversing it and taking the complement (*i.e.* swapping `C` with `G` and `A` with `T`) each gene is encoded by a single strand—the *coding strand* for that gene. The first step required to activate a gene is the binding of the protein complex RNA Polymerase to its *promoter*

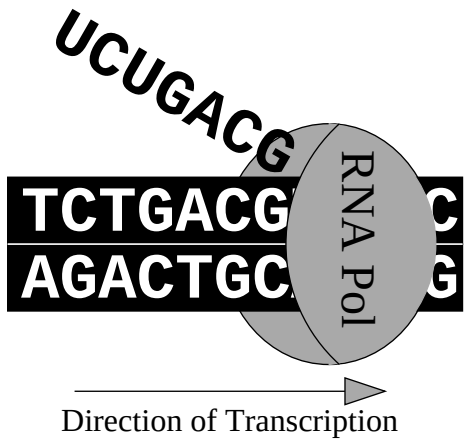


Figure 1-2: RNA Polymerase, DNA (white letters on black), and RNA (black letters) form a ternary complex during transcription.

region. This is the region of DNA at the beginning (5′-most end) of the gene, and it tells the cell under what conditions the gene should be expressed, and also on which strand the gene lies. RNA Polymerase makes an exact RNA copy of one strand. RNA is similar to DNA but uses a different backbone, typically has only one strand, and uses the chemical letter uracil (abbreviated `U`) instead of `T`. The process of copying a gene into an RNA is called *transcription*.

The RNA is transcribed from its 5′ end to its 3′ end, so that RNA Polymerase, DNA and the growing RNA strand form a ternary complex at any point during transcription (Figure 1-2). The transcript ends after recognition of a cleavage signal sequence by protein factors associated with RNA Polymerase.

A complete RNA transcript is also known as a pre-messenger RNA, or *pre-mRNA*, and is an exact RNA copy of a contiguous region of one strand of the DNA genome. It is at this point that the splicing reaction takes place (see section 1.2 below), and introns are excised, leaving the mature messenger RNA, or *mRNA*. Two other changes associated with the maturation of the pre-mRNA are the addition of a special nucleotide known as a *cap* to the 5′ end of the molecule, and the addition of a 3′ tail of typically 100-200 adenosines.

This capped, spliced and polyadenylated mRNA is then exported from the nucleus into the cytoplasm. In the cytoplasm it is bound by ribosomes, large complexes of RNA and protein which catalyze the translation of the mRNA into protein. The

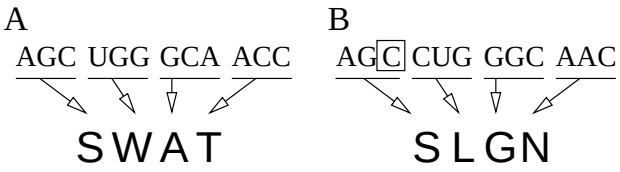


Figure 1-3: The genetic code is highly sensitive to nucleotide insertion and deletion. (a) Proteins are made by translating consecutive RNA nucleotide triplets into the 20-letter amino acid alphabet. RNA sequence is shown in small letters and protein sequence in big letters. (b) A single nucleotide insertion C in the RNA sequence completely changes the rest of the protein sequence.

ribosome begins translation near the 5′ end of the mRNA, and reads triplets of nucleotides, or *codons* (Figure 1-3a). Each codon corresponds to a single amino acid from the 20 amino acid alphabet. Three of the sixty-four (4^3) codons are *stop codons* and signal the ribosome to terminate and release the growing protein. A given stretch of RNA could in principle encode three different proteins, according to whether the first codon begins with the first, second, or third nucleotide. Therefore, the insertion or skipping of even a single nucleotide at any point in the mRNA will render the rest the of the protein nonsense (Figure 1-3b). Thus even the slightest error in the production of the mRNA can have a profound effect on the production of a functional protein.

1.2 Splicing the Interrupted Gene

Splicing was first discovered in viruses in 1977 [3, 8]. The investigators were studying which parts of the DNA were being transcribed into RNA. They were able to isolate a single type of mRNA from the virus, and also isolate the non-coding or *template* strand of DNA from which the mRNA was expressed. When single-stranded DNA and complementary RNA are mixed they form a DNA:RNA hybrid, which is a double-stranded polymer with similar railroad-type structure and base-pairing properties as native DNA. When this single type of mRNA was hybridized to the single template strand of DNA, electron micrographs revealed that there were regions of the DNA that would hybridize with the mRNA interrupted by other regions that would loop out without forming base pairs (Figure 1-4).

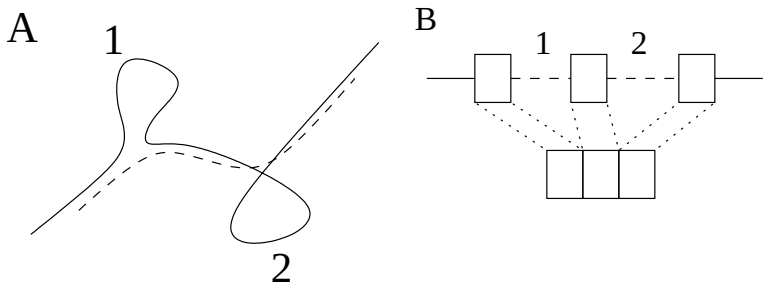


Figure 1-4: The discovery of interrupted genes. (A) Schematic of electron micrograph of single-stranded DNA (solid line) hybridized to mRNA (dashed line). The loops in the DNA represent introns which, absent in the mRNA, remain unbound. (B) Diagram of a three-exon/two-intron gene. Exons, boxes. Non-transcribed DNA, solid lines. Introns, dashed lines. Top, pre-mRNA. Bottom, mature mRNA. Dotted lines indicate correspondence of splice junctions in pre-mRNA and mature mRNA.

The entire region was transcribed, but the regions that looped out without forming base pairs, later termed *introns*, were then excised from the pre-mRNA. The regions that remained in the mature mRNA, termed *exons*, were able to base pair with the corresponding regions from the DNA.

The idea that genes were not contiguously encoded in the genome came as a shock, and a great deal of energy was put into understanding both the mechanism by which the splicing reaction took place and also the manner in which the excisions were specified in the RNA sequence. Although much progress has been made, both problems remain at least partially open today. The *spliceosome*, the cellular machinery responsible for the splicing reaction, is one of the largest and most complex of the entire cell. Although most of its components are known, many of their roles in the context of the larger machine are still poorly understood. As for the elements in the RNA sequence that specify the ends of the introns, they have been described very well in yeast [56], but remain only partly understood in humans and other animals.

As more and more introns and exons were mapped, it became clear that there were at least three such elements in every splicing reaction. Two of three are the $5'$ and $3'$ splice sites, located respectively at the beginning and end of every intron (Figures 1-5 and 1-6). They act as flags that say "cut here" to the spliceosome, like the lines drawn by a carpenter on a wooden beam to guide the saw. The third element, known as the *branch site*, or, when it needs to be abbreviated, the *branch point site* (BPS), is

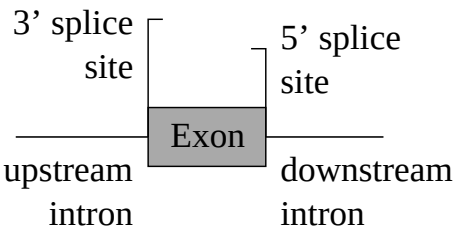


Figure 1-5: The parts of an exon. Every exon (gray box) begins with a 3′ splice site (right-facing periscope) and ends with a 5′ splice site (left-facing periscope).

Figure 1-6: Human splice sites. The 3′ splice site (left) appears upstream and the 5′ splice site (right) appears downstream of every exon. The height of each letter at each position is proportional to the number of actual human splice sites that have that letter in their DNA sequence at that position relative to the splice junction.

located usually in the last 20-30 bases of the intron, near the 3′ splice site. It turned out that this functioned as a sort of handle by which the spliceosome could grab the excised intron, and also as a marker to tell the spliceosome to look nearby for the 3′ splice site.

Another idea that emerged from research into human RNA splicing was that the basic unit of recognition is the exon, not the intron [46]. Rather than looking at the pre-mRNA as mostly exon with a few introns that need to be excised (as in yeast), the human spliceosome needs to identify and join the short exons from a sea of introns: exons tend to be about 125 nucleotides long, with very few any longer than 250 nucleotides, whereas human introns tend to be about 10 times longer.

This “exon definition” model, then, suggested that the spliceosome searches pre-mRNA for a branch site followed very soon by a 3′ splice site and then by a 5′ splice

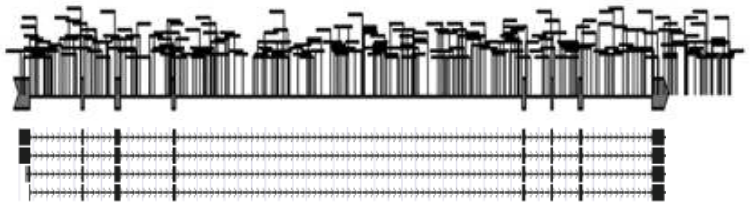


Figure 1-7: Human Sirt1 gene splice sites. (Top) Sequences along the DNA are scored for their similarity to the $3'$ splice site (right-facing periscope) and $5'$ splice site (left-facing periscope) consensus sequences (Figure 1-6). The heights of the periscopes are proportional to the sequence scores, and are plotted only for those scores above a minimum threshold. (Bottom) Known Sirt1 mRNAs from GenBank (see section 1.5). Black boxes indicate mRNA sequence (exons) and cross-hashed lines indicate intronic or genomic sequence.

site at a distance of at most another 250 nucleotides. Surprisingly, if we look at the distribution of sequences that are similar to consensus in Figure 1-6 we see that they are abundant in a typical human gene. As an example, Figure 1-7 shows the extent to which such sequences, known as *decoy splice sites*, pepper the length of the human Sirt1 gene, and the extent to which most of them are nevertheless ignored. The cell always seems to include the correct sequences as exons into the mature mRNA and exclude the rest.

Such a picture is typical for many human genes, and suggests two basic questions about splicing. First, how is it that the inclusion of true exons into the transcript is enforced, and second, how is it that pairs of nearby splice sites with good consensus scores (*pseudoexons*) are consistently excluded from mature mRNA.

Part of the answer to these questions lies in short sequences called *exonic splicing enhancers* (ESEs). These sequences, which are present in nearly all human exons, serve as binding sites for members of the SR protein family. SR proteins recruit components of the spliceosome to the upstream $3'$ and downstream $5'$ splice sites, thereby defining ESE-containing pre-mRNA segments as exons [20, 62, 15]. There has also been a report that the SR proteins, when bound to ESEs, suppress the skipping of that exon, so that an upstream $5'$ splice site cannot be joined with a downstream $3'$ splice site when they are separated by ESEs [26].

In fact, another class of sequences, the *exonic splicing silencers* (ESSs) can suppress the inclusion of a putative exon in which they are located [61]. These elements

are abundant in pseudoexons, which is one mechanism by which they are consistently excluded. *Intronic splicing enhancers* (ISEs) and *silencers* (ISSs) have also been studied to some extent. These are elements that appear in introns nearby exons or pseudoexons and function to enhance or repress their inclusion in the mRNA.

Chapter 3 describes our discovery of a new class of putative splicing regulatory sequences. These are *collocations*, or pairs of elements which function in concert with each other. In one particularly interesting example, the GC/AU collocation, a G+C-rich motif appears at the beginning of introns paired with an A+U-rich motif at their end. We believe that this pair of motifs cooperatively represses the inclusion of pseudoexons throughout the length of the intron.

1.3 Alternative Splicing

It is possible for cells to modulate gene expression to their advantage. As an example consider the gene AQP in frogs. This gene, which has a human homolog, encodes the protein aquaporin, which resides in cellular membranes and makes them water-permeable. Without the expression of such proteins we would not, for example, be able to take in water from our environment and would surely die. On the other hand, aquaporins are absent in frog egg cells. When, by artificial methods, such cells are forced to express aquaporins and then placed in water with low salt concentration, such as would be found in a fresh water pond, they burst. This is due to the osmotic influx of water from low to high salt concentration. Thus the regulation of the expression of AQP is essential for the life of a frog: off in eggs, and then on in various adult cells.

It seems that every step along the gene pathway is regulated at some point by some part of the human body. At the most basic level, we know that the initiation of transcription is tightly controlled, with most genes completely shut off in most cells. Once a gene is transcribed into an mRNA, its export to the cytoplasm could be regulated. Once in the cytoplasm, its half-life can be regulated. Next, the various steps of translation can be regulated. Finally, the half-life of the translated protein can

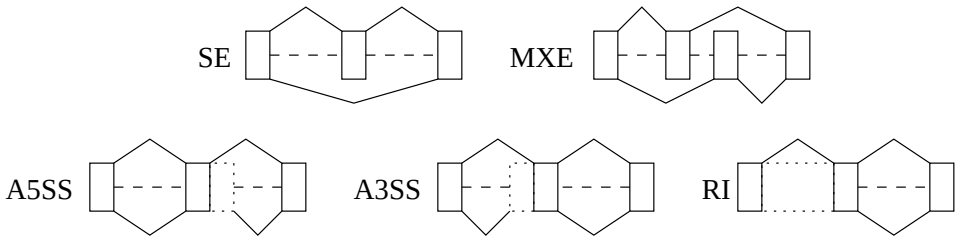


Figure 1-8: Five common patterns of alternative splicing. Exons are shown as boxes, introns or genomic sequence as dashed line, splicing pathways as solid lines. (SE) Skipped exon. One pathway (upper) includes the middle exon but the other skips it. (MXE) Mutually exclusive exon. Each pathway includes one or the other, but not both, of the two middle exons. This is exemplified by α-actinin. (A5SS) Alternative $5'$ splice site. One pathway (upper) uses the first $5'$ splice site and the other uses the second. Thus the upper pathway gives rise to a shorter transcript than the lower. (A3SS) Alternative $3'$ splice site. Similar to A5SS, but at the $3'$ splice site. (RI) Retained intron. One pathway (upper) splices out the first intron, but the other includes it in the mRNA, creating a longer transcript.

be regulated. Each of these steps, when positively regulated, lead to greater expression of the encoded protein, and when negatively regulated lead to lower expression or the complete shutting off of the protein.

When the splicing of pre-mRNA to mRNA is regulated the effect can be much more subtle. In one case, since it is known that some genes cannot be efficiently exported until they are spliced, if the rate of splicing is regulated, it will simply affect the final level of the protein. Surprisingly, there are many examples where the actual choice of splice sites is regulated, so that it is not only the final protein level but the actual sequence of the protein that is regulated! This is called *alternative splicing*, and the different mature mRNAs that can be spliced out of a given pre-mRNA are known as *isoforms*. It has been estimated that at least half of all human genes can be alternatively spliced [27]. Some of the common patterns of alternative splicing are shown in Figure 1-8.

An interesting example of alternative splicing is the gene α-actinin, which functions in the cytoskeleton, the intracellular matrix that gives order and shape to every cell. This gene contains among its exons two in a row that are known as NM and SM. These exons are *mutually exclusive*—the spliceosome always chooses one or the other, but not both, to include in the mature mRNA. These exons are regulated in such a fashion that smooth muscle cells generally produce the SM-containing isoform,

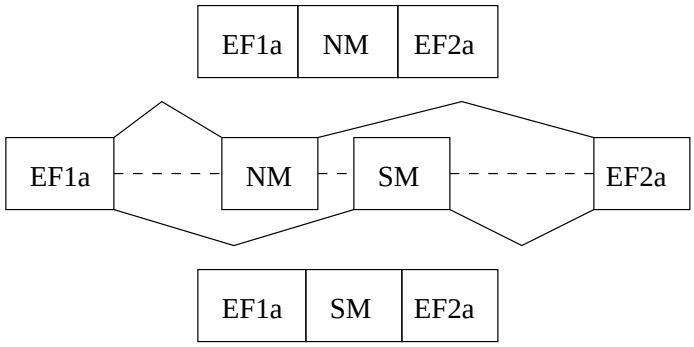


Figure 1-9: Alternative splicing of *alpha*-actinin. (Top) NM-containing isoform spliced in non-muscle cells. (Middle) *alpha*-actinin pre-mRNA with two splicing pathways. Dashed lines represent introns and solid lines represent splices. (Bottom) SM-containing isoform splice in smooth muscle cells. (Figure modified from [58].)

whereas non-muscle cells generally produce the NM-containing isoform (see Figure 1-9).

Alternative splicing is important because it multiplies the diversity of proteins that can be expressed in one organism. For example, the fly Dscam gene can be alternatively spliced in over 38,000 ways [50]. In this extreme example, a single gene can in theory give rise to more variant proteins than there are genes in the entire genome. Sometimes the variants of the protein will have different properties in the cell, and sometimes the changes are more important at the RNA level—for example one splice might include a sequence that targets the mRNA to be degraded. Therefore, understanding alternative splicing is key to a complete picture of gene expression.

Our interest was in probing the role of the branch point in alternative splicing which has, to this point, been poorly studied. We were motivated by the α-actinin exons, whose mutual exclusion, we shall see, is due to the position of their branch points. Chapter 4 describes our creation of the largest known set of branch points in human and mouse, and what we might learn from examining them. One interesting discovery is that the ordering of otherwise normal regulatory elements can lead to the skipping of an exon.

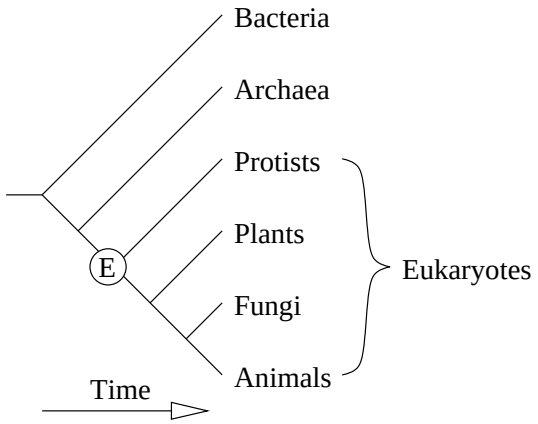


Figure 1-10: The three domains of life. Eukaryotes are the organisms with nuclei, and also the only ones known to date that have spliceosomes. They include all plants, animals, fungi and the early-branching protists. The 'E' within a circle denotes the last common ancestor of all eukaryotes.

1.4 The Evolution of Introns

It was not long after the discovery of splicing that researchers began to wonder about their evolutionary origins. Why weren't genes encoded contiguously? How widespread was this phenomenon in the tree of life? What advantage would cells with introns have over those without?

In any evolutionary study the first task is to tell the history, and only then can one begin to speculate the reasons. In the case of splicing, it was crucial to know exactly when introns arose in the history of life, and then how quickly and where they spread.

To date, no bacterium or archaeon has been found that has a spliceosome, but every eukaryote studied does (see Figure 1-10), from the most complicated multi-cellular animals and plants all the way down to simple single-celled organisms such as yeast, *Giardia* and *Plasmodium* that do not need alternative splicing to diversify their protein production. Therefore it is very likely that the last ancestor of the eukaryotes had a spliceosome.

This suggests two more questions: How much earlier before that ancestor was the spliceosome invented, and how many introns did that ancestor have? These questions have fueled an intense debate. On the one hand, the *introns early* camp maintained

that the very earliest life had introns, and that they were lost along the bacterial and archaeal lineages. This camp also put forth the *exon theory of genes* which said that most modern genes today arose from shuffling and recombination of ancient exons. They also propose that the eukaryotic ancestor had many introns, much like mammals and other vertebrates today. Yeast, which has much fewer, they maintain came about through the loss of introns. The *introns late* camp proposed that the spliceosome arose in an early eukaryotic cell and furthermore that most introns that we see today arose after the split of the plant and opisthokont (animal + fungal) lineages.

In recent years, with wide-scale availability of complete genome sequences and exon annotations, it is finally possible to resolve at least the question of intron number using phylogenetic methods (see 1.6.1). Chapter 2 shows our effort to mathematically model the gain and loss of introns in four fungi. This represents one of the earliest and most thorough examinations of genome-scale patterns of intron gain and loss. We find that, at least in this clade of fungi, gain and loss are roughly in balance. Their common ancestor had approximately the same number of introns as they do today, and that number has persisted.

1.5 Biological Data

The field of computational biology has arisen over the past few decades due to the abundance of data that has been generated from a variety of high-throughput technologies. These data include nucleic acid sequence data, microarray data, and a variety of protein data. The nucleic acid sequence data, with which we are primarily concerned here, includes genomic as well as transcript sequence. Microarrays are a method of measuring in parallel the relative abundances of thousands or even millions of different species of nucleic acid by applying cell extract to a chip with probes complementary to sequences of interest. Protein data include sequences of purified proteins by Edman degradation or mass spectrometry, measurements of protein abundances by antibody arrays, and three-dimensional structures solved by X-ray

diffraction or NMR spectroscopy. For the work contained in this thesis, the most important data are nucleic acid sequences.

Of the nucleic acid sequences, the most fundamental are the genomic. Entire genomes can be sequenced by isolating DNA from cells, shearing the DNA into shorter pieces, and then sequencing these short pieces. This method is called shotgun sequencing and requires the subsequent computational assembly of the different pieces. Genomic sequences can be downloaded from a number of websites. Perhaps the most important are GenBank (http://www.ncbi.nlm.nih.gov/), Ensembl (http://www.ensembl.org), and UCSC (http://genome.ucsc.edu/).

The other important class of sequence data that we use are transcript data. These are sequences of RNA from cells and are generally of two types: full-length mRNA and EST. Full-length mRNAs traditionally represent the efforts of individual research groups in sequencing genes of interest. mRNAs differ from the genomic sequence in that introns are absent. By aligning the mRNA sequence to the genome one can tell where transcription starts, where cleavage (and polyadenylation) occurs, and where each exon begins and ends. EST (expressed sequence tags) are from high-throughput sequencing projects. An EST project begins with the aggregation of RNA from a large number of cells. These cells will generally be similar, for example all from the same organ or cell line. Short pieces of RNA are then amplified and sequenced, so that each sequence represents an internal region of a longer mRNA. These short sequences are called ESTs. When they are aligned against the genome they tell us not only that the gene in that part of the genome is expressed in the corresponding cell, but also allow exon boundaries to be defined if they happen to span an intron (see Figure 1-11).

Collectively, ESTs give us the most complete picture available to date of the variety of human splicing isoforms. Currently there are 7.7 million human ESTs. With about 25,000 genes that gives us a coverage of roughly 300 ESTs per gene. Most ESTs overlap at least one intron, and so by examining the boundaries of their alignments to the genome we can create a database of all known human splice sites. By analyzing the sequence near these splice sites we can learn about what determines

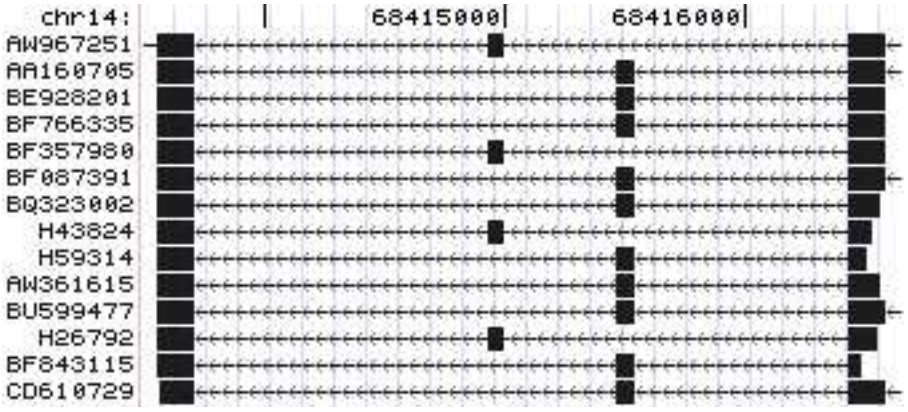


Figure 1-11: Exons 20-24 of the alpha-actinin gene. Coordinates along chromosome 14 are shown at the top, and a sample of EST sequences (ID codes on left) aligned back to the genome are displayed below. The black boxes represent exons, or aligned sequences, and the thin lines represent introns. The small left-pointing arrows indicate that the gene is on the minus strand of the chromosome. The middle two exons, 21 and 22, can be seen to be mutually exclusive; the two are never seen together in the same transcript. Different boundaries of exons 20 and 24 reflect the beginning and end of ESTs. Image modified from `http://genome.ucsc.edu` [28].

patterns of splicing in human cells.

1.6 Algorithms of Computational Biology

1.6.1 Phylogeny

Phylogeny is the re-creation of the tree of life from characteristics of modern organisms. In the past, these characteristics (or "characters") were typically morphological: number of appendages, typical size of a species, type of skin, etc. Today we can use sequence analysis to infer evolutionary history. We assume only that sequences that are similar must be evolutionarily related, and that the time since divergence of dissimilar sequences must be longer.

We often model *phylogenetic trees* as rooted binary trees. Each leaf represents a modern organism, and each internal (non-leaf) node represents a hypothetical ancient organism. In particular, the root node represents the last common ancestor of all the modern organisms. Each bifurcation in the tree represents a *speciation event*—one species becoming two. For example, in Figure 1-10, we imagine that there was a

species of single celled organisms that was the progenitor of all modern eukaryotes. At the point marked E in the figure, this population somehow split into two groups, one becoming the progenitor of all modern plants, and the other of modern fungi and animals.

The simplest way to construct a phylogenetic tree from a set of genes is called the Unweighted Pair Group Method with Arithmetic Mean (UPGMA). This method starts with a symmetric matrix D whose rows and columns are indexed by the set of modern organisms and whose entries are the evolutionary distances between the corresponding pairs of organisms. Such a matrix could be constructed, for example, by looking at percent mismatch in alignments of homologs of the same gene in all organisms, or by taking the average of pairwise percent mismatch over a large number of genes.

The UPGMA tree is constructed recursively, from the leaves up to the root. At each step the minimum entry d_{ij} in D is selected. In other words, we choose the pair of organisms i and j that are most closely related. These two are connected with an internal node. Then i and j are removed from D and a row and column corresponding to the internal node (say $\overline{ij}$) is added. The new distances are calculated by the arithmetic mean, or $d_{\overline{ij},k} := (d_{ik} + d_{jk})/2$, and the process is iterated until all the nodes have been added to the tree.

We use the UPGMA method in chapter 2 to create the tree relating four fungi.

1.6.2 Motif Finding Using k-mer Statistics

The most naïve, and yet powerful, type of sequence analysis is the study of k-mer statistics. A k-mer is a word of length k from a fixed alphabet (in our case usually the DNA alphabet $\Sigma = \{\texttt{A},\texttt{C},\texttt{G},\texttt{T}\}$). Typically an investigator starts with a set of sequences B known to have a certain biological function, and then some control sequences C, and, after choosing a fixed length k, for each $x \in \Sigma^k$ tallies some measure of the abundance of x in the sequences of B relative to its abundance in C. Those x enriched in B are then predicted to have some biological activity.

One elegant use of k-mer statistics has been the identification of exonic splicing

enhancers [15]. In this case the investigators were interested in finding exonic splicing enhancers (ESEs). These are sequences that appear within exons that promote splicing. They are thought to function by recruiting multi-domain proteins that can both bind the ESE and stabilize the spliceosome at the nearby splice sites.

The investigators began with a set of constitutive exons E. They created this set by looking for exons with EST support that had no evidence of alternative splicing (e.g. any of the classes in Figure 1-8). They then scored the splice sites of the exons according to their similarity with the consensus sequences in Figure 1-6, and separated the exons into those with strong splice sites (close to consensus) E_s or weak splice sites (differing from consensus) E_w. They also created a set of introns I that, according to EST evidence, were never known to be included in any transcript. In this study they examined hexamers (i.e. $k = 6$). For each hexamer x they then calculated its frequency in the four different sets E, E_s, E_w and I (that is, the number of times x is seen in that set divided by the total number of hexamers in the set).

They next considered that a hexamer that functioned to enhance splicing in exons should satisfy two properties. First, it should be more prevalent in exons E than introns I, because if it were to occur in introns it might recruit the spliceosome to the wrong place. Second, it should be more prevalent in exons with weak splice sites E_w than exons with strong splice sites E_s, because it is precisely those with weak splice sites that need "the extra help", and therefore should experience stronger selection to conserve ESEs. Therefore they looked for hexamers for which both f_E/f_I and f_{E_w}/f_{E_s} were above some threshold. Indeed when they tested such hexamers by inserted them into a mini-gene construct where the middle exon is typically skipped they found that they could enhance their inclusion, proving their role as ESEs.

In chapter 3 we develop a new k-mer statistical method for detecting pairs of sequences that interact at a distance.

1.6.3 Hidden Markov Models

Hidden Markov Models (HMMs) are a more complicated way of analyzing sequence data that seeks to annotate entire regions of sequence. Gentler introductions than

we present here can be found in the literature (e.g. [14]). A (discrete) HMM is a probability distribution on the set of pairs of sequences of equal length, the first taken from the *alphabet of hidden states* S, and the second from the *emitted alphabet* Σ. These can be any finite sets, but for our purposes we think of Σ as either protein or DNA sequence, and of S as some useful annotations for different parts of a gene, such as "exon" and "intron".

The HMM is defined by a *transition matrix* $T \in \mathbb{R}^{S \times S}$, and an *emission matrix* $E \in \mathbb{R}^{S \times \Sigma}$. The marginal probability of the state sequences are just Markov probabilities taken from T. In other words, (ignoring the probability of the starting state) $P(\underline{s_1 \cdots s_n}) = \prod_{i=2}^{n} t_{s_{i-1} s_i}$. Then, conditional on a given state sequence, the probability of the emitted sequence is given by multiplying independently the emission probabilities for corresponding positions in the sequences: $P(\underline{\sigma_1 \cdots \sigma_n} | \underline{s_1 \cdots s_n}) = \prod_{i=1}^{n} e_{s_i \sigma_i}$.

Given a sequence $\underline{\sigma}$ there are efficient algorithms to (1) calculate its marginal probability, (2) find the *parse*, or sequence of hidden states $\underline{s_1 \cdots s_n}$, that emits it with maximum likelihood, (3) to calculate the probabilities, conditional on $\underline{\sigma}$, of any particular state at a particular position, and (4) to find the maximum likelihood estimates of all the entries in T and E.

The *forward algorithm* gives the probability of the sequence $\underline{\sigma}$, say of length n. It uses dynamic programming to fill the matrix F, which represents probabilities of emission of initial segments of $\underline{\sigma}$ ending in particular states: $f_{is} := P(\underline{\sigma_1 \cdots \sigma_i}, s_i = s)$. f_{is} satisfies the recursion

$$f_{is} = \sum_{s' \in S} f_{i-1,s'} t_{s's} e_{s\sigma_i}.$$

The total probability is then $\sum_{s \in S} f_{ns}$. Similarly, the *backwards algorithm* fills the matrix B whose entries are defined as $b_{is} := P(\underline{\sigma_i \cdots \sigma_n} | s_i = s)$ for all $1 \leq i \leq n$ and $s \in S$, and gives the total probability $\sum_{s \in S} b_{1s}$, which is identically $\sum_{s \in S} f_{ns}$.

The *Viterbi algorithm* calculates the maximum likelihood parse given $\underline{\sigma}$. It uses a

similar recursion, replacing the sum with a maximum:

$$P_{\mathrm{ML}}(\underline{\sigma_1 \cdots \sigma_i}, s_i = s) = \max_{s' \in S} P(\underline{\sigma_1 \cdots \sigma_{i-1}}, s_{i-1} = s') t_{s's} e_{s\sigma_i}.$$

Typically the calculation is done with the logarithms of these probabilities:

$$\log P_{\mathrm{ML}}(\underline{\sigma_1 \cdots \sigma_i}, s_i = s) = \max_{s' \in S} \left[\log P(\underline{\sigma_1 \cdots \sigma_{i-1}}, s_{i-1} = s') + \log t_{s's} + \log e_{s\sigma_i} \right].$$

Thus the maximum likelihood parse of the entire sequence has probability $P_{\mathrm{ML}}(\underline{\sigma}) = \max_{s \in S} P_{\mathrm{ML}}(\underline{\sigma}, s_n = s)$. Keeping track of the states that give the maxima at each step of the algorithm allows the entire parse to be reconstructed.

The probabilities in F and B can be used to calculate *posterior probabilities* for a state s given the sequence $\underline{\sigma}$. In other words, for each position i we can calculate $P_s^{\mathrm{post}}(i) := P(s_i = s, \underline{\sigma})$. This is done by multiplying the forward and backward probabilities together:

$$P_s^{\mathrm{post}}(i) = \frac{f_{is} b_{is}}{e_{s\sigma_i}}.$$

Dividing by $P(\underline{\sigma})$ turns this into a conditional probability, so that it can be interpreted as the probability that position i was emitted by state s.

Finally, training the HMM is accomplished by an expectation-maximization (EM) procedure known as the *Baum-Welch algorithm*. Given $\underline{\sigma}$ we calculate the maximum likelihood estimates of E using

$$\widehat{e_{s\sigma}} := \frac{\sum_{i:\sigma_i = \sigma} P_s^{\mathrm{post}}(i)}{\sum_{i=1}^{n} P_s^{\mathrm{post}}(i)} \tag{1.1}$$

The numerator is the number of times we see state s emitting letter σ, averaged over all parses of $\underline{\sigma}$, and the denominator is the number of times we see state s emitting any letter, also averaged over all parses. For T the calculation is only slightly more complicated. To get $\widehat{t_{ss'}}$ we would like to have a term that represents the number of times we see state s followed by state s' in the hidden state sequence. We will divide

this by a term that represents the number of times we see state s at all, being careful about boundaries:

$$\widehat{t_{ss'}} := \frac{\sum_{i=1}^{n-1} f_{is} t_{ss'} b_{i+1,s}}{\sum_{i=1}^{n-1} P_s^{\text{post}}(i)}. \tag{1.2}$$

The forward-backward algorithm, then, is to cycle through calculating F and B, then updating E and T according to equations 1.1 and 1.2.

There are two simple but useful variants of the Baum-Welch algorithm. The first variant uses multiple sequences for training. The only change is to extend the sums in Equations 1.1 and 1.2 to sum over all the sequences. The other variant is known as *Viterbi training*. Here the numerators and denominators in these equations, which represent expected counts, are replaced by the actual counts taken from maximum likelihood parses of the sequences. The name Viterbi Training highlights that the maximum likelihood parse generated by the Viterbi algorithm is used to update these parameters.

HMMs and an important generalization, the *hidden semi-Markov models* have had an important role in sequence analysis. In chapter 4 we apply them to the problem of identifying branch points in introns to predict the largest set of high-confidence branch points to date.

Chapter 2

Patterns of Intron Gain and Loss in Fungi

Abstract

Little is known about the patterns of intron gain and loss or the relative contributions of these two processes to gene evolution. To investigate the dynamics of intron evolution, we analyzed orthologous genes from four filamentous fungal genomes and determined the pattern of intron conservation. We developed a probabilistic model to estimate the most likely rates of intron gain and loss giving rise to these observed conservation patterns. Our data reveal the surprising importance of intron gain. Between about 150 and 250 gains and between 150 and 350 losses were inferred in each lineage. We discuss one gene in particular (encoding 1-phosphoribosyl-5-pyrophosphate synthetase) that displays an unusually high rate of intron gain in multiple lineages. It has been recognized that introns are biased towards the 5′ ends of genes in intron-poor genomes but are evenly distributed in intron-rich genomes. Current models attribute this bias to 3′ intron loss through a polyadenosine-primed reverse transcription mechanism. Contrary to standard models, we find no increased frequency of intron loss toward the 3′ ends of genes. Thus, recent intron dynamics do not support a model whereby 5′ intron positional bias is generated solely by 3′-biased intron loss.

2.1 Introduction

Over a quarter of a century after the discovery of introns, fundamental questions about their function and evolutionary origins remain unanswered. Although intron density differs radically between organisms, the mechanisms by which introns are inserted and deleted from gene loci are not well understood. A correlation has been observed between intron density and positional bias [35]. Introns are evenly distributed within the coding sequence of genes in intron-rich organisms, but are biased toward the 5′ ends of genes in intron-poor organisms. This bias is particularly pronounced in the yeast *Saccharomyces cerevisiae*. It has been suggested that both the paucity and positional bias of introns in yeast may be due to intron loss through a mechanism of homologous recombination of spliced messages reverse-transcribed from the 3′ polyadenylated tail [19]. This reverse transcription mechanism was first demonstrated in experiments with intron-containing Ty elements in yeast [4]. More recently, [35] concluded that homologous recombination of cDNAs is the simplest explanation for the positional bias observed in all intron-poor eukaryotes. However, few data exist concerning the actual mechanisms and dynamics of intron evolution.

Fungal genomes are in many ways ideal for exploring questions of intron evolution. The fundamental aspects of intron biology are shared between fungi and other eukaryotes, making fungi appropriate model organisms for intron study. They are gene dense with relatively simple gene structures compared with plants and animals, making gene prediction more accurate. Fungi also display a wide diversity of gene structures, ranging from far less than one intron per gene for *S. cerevisiae*, to approximately 1-2 introns per gene on average for many recently sequenced ascomycetes (including the organisms in this study), to roughly seven introns per gene on average for some basidiomycetes (e.g., *Cryptococcus*). Finally, fungi display a strong 5′ bias in intron positions, enabling us to investigate the processes underlying this phenomenon.

In principle, a 5′ intron bias could arise through various combinations of intron gain and loss, and a complete understanding of intron positional bias requires an assessment of the contributions of both of these processes. A number of studies demon-

strate the occurrence of intron gain and loss in individual genes or gene families. [34] offered early examples of well-supported intron gain by comparing triose-phosphate isomerase genes from diverse eukaryotes and demonstrated that numerous introns could be most parsimoniously explained by a single gain with no subsequent losses. [41] later provided evidence for de novo intron insertion into the otherwise intron-less mammalian sex-determining gene SRY. Evidence for the occurrence of multiple independent intron losses has also been reported in studies such as [47] which inferred gain and loss events in a family of chemoreceptors in Caenorhabditis elegans.

More recently, a number of genome-wide studies of intron dynamics have been conducted. [49] described genome-wide comparisons between human and mouse (with Fugu as an outgroup) and between mouse and rat (with human as an outgroup), and observed a sparseness of intron loss and complete absence of intron gain in these closely related organisms. On the other hand, [48] observed an abundance of lineage-specific intron loss and gain when analyzing clusters of orthologous genes in deeply branching eukaryotes. Similarly, [43] analyzed ten protein families in distantly related eukaryotes, with a single prokaryotic outgroup, and obtained evidence that extant introns are predominantly the result of intron gains. In search of clues to understand the mechanism of intron gain, [18] aligned introns from various eukaryotes, and [10] applied a similar approach in a comparative study of nematodes. None of these studies addressed the positional bias of intron gain and loss events. Here we report the results of a genome-wide comparative analysis of intron evolution in organisms that have a strong $5'$ bias in intron location and are at an appropriate evolutionary distance to reveal positional trends in intron gain and loss.

2.2 Results

2.2.1 Construction of a Set of 2,000 Orthologs

To investigate the roles of both gain and loss in intron evolution, we compared the genomes of four recently sequenced fungi spanning at least 330 million years of evolu-

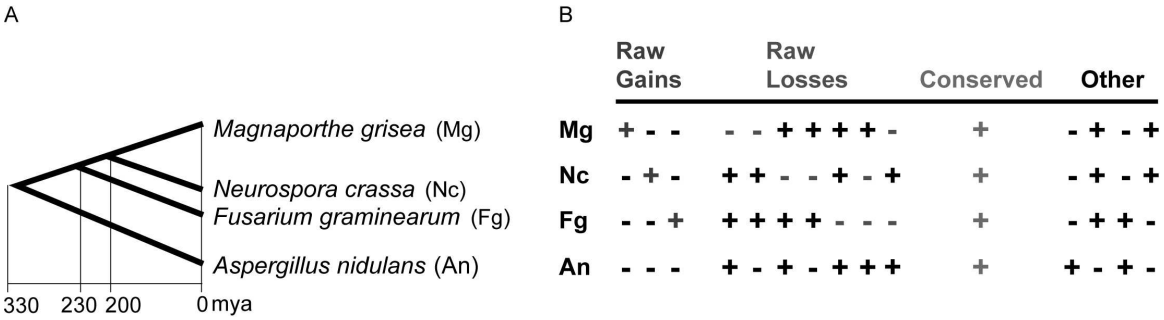


Figure 2-1: Phylogenetic Tree and Intron Conservation Patterns. A, Phylogenetic tree of the four fungal organisms studied (*M. grisea*, *N. crassa*, *F. graminearum*, and *A. nidulans*) with estimated time scale in millions of years ago. The rooted organismal tree was constructed using an unweighted pair group method using arithmetic averages based on a concatenated alignment of 2,073 orthologous gene sets. Estimated dates of divergence from [60], [2] and [23]. B, Classification of intron presence (+) and absence (—) patterns across the four fungal species. A blue "+" indicates a raw intron gain in the corresponding organism, a red "—" indicates a raw intron loss in the corresponding organism, a green "+" indicates a conserved intron, and all other introns are indicated in black.

tion [60], [2], [23] (Figure 2-1): *Aspergillus nidulans*, *Fusarium graminearum*, *Magnaporthe grisea*, and *Neurospora crassa*. Ortholog sets composed of one gene from each of the four genomes were identified as pairwise best bidirectional BLAST hits satisfying stringent overlap criteria. Orthologs in each set were subsequently aligned, and the locations of introns were marked. These intron positions (regions of the multiple sequence alignment containing an intron in at least one of the four sequences) were subjected to rigorous alignment quality filtering to eliminate alignment and annotation errors (Figure 2-2A). To set the filtering thresholds, we manually classified ten residue alignment windows on either side of 181 randomly selected intron positions as "clearly homologous," "possibly homologous," or "non-homologous." Requiring 30% identity and 50% similarity in these windows captured 92% of the clearly homologous positions, 29% of the possibly homologous positions, and only 2% of the non-homologous positions (Figure 2-2B). Passing intron positions were split into five quintiles according to their relative position within the annotated coding sequence.

2.2.2 Genome-Wide Characterization of Intron Conservation

We applied our analysis protocol to 2,073 putative ortholog sets that included 9,352 intron positions. Of these initial intron positions, 5,811 were removed because of low

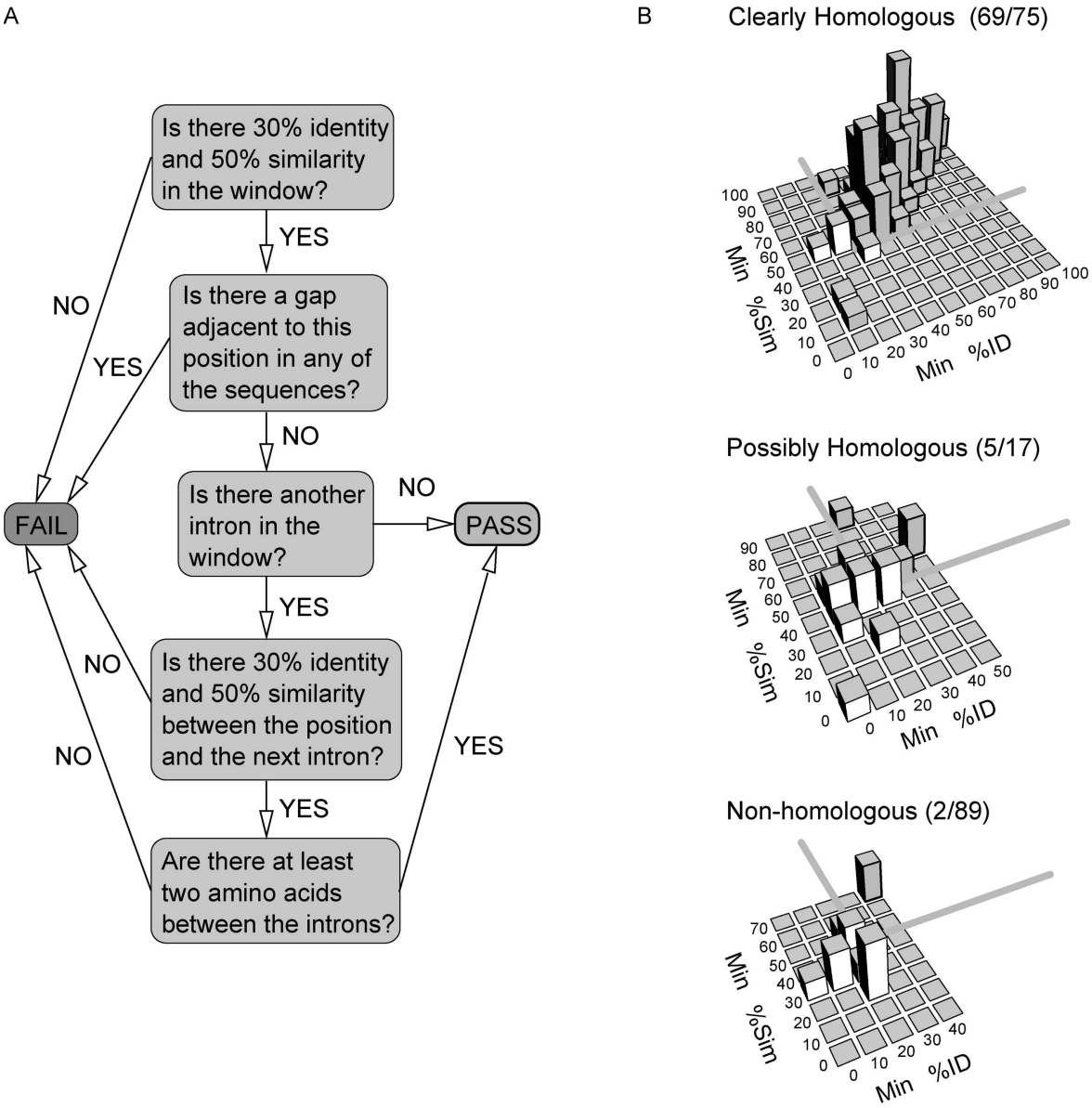


Figure 2-2: Alignment Filtering Protocol. A, Schematic of filtering protocol applied to a ten-residue window on each side of every intron position. If either side failed the filter, the position was discarded. B, Distributions of minimum percent identity and similarity in ten-residue windows around 181 randomly selected intron positions, for three manual classifications. The minima were taken between the left and right windows. The yellow lines indicate the chosen thresholds of at least 50% similarity and 30% identity, and bars are colored yellow if they fall above the thresholds (pass) or orange if they fall below the thresholds (fail). Parentheses indicate the number of introns in each class that pass the cutoff and the total number of introns in that class. The five lowest-percent identity and similarity bars, containing 77 positions, in the "non-homologous" plot are omitted so as to not obscure the rest of the histogram.

conservation surrounding the intron, or because of an adjacent gap, or both. It is possible that some of the positions neighboring gaps may in fact reflect intron gain or loss events that occurred simultaneously with coding sequence insertion or deletion [32]. However, removing these positions did not significantly impact our results, as the number of positions adjacent to gaps was only about one-tenth of the number of positions that passed the quality filter, and the removal of these introns did not alter the apparent positional bias of the overall distribution (Figure 2-3). An additional 92 introns had nearby introns with insufficient conservation between the two introns and were thus also rejected.

In the end, a total of 3,450 intron positions (roughly 37% of intron positions considered) passed the quality filter. The complete set of aligned orthologs with passing and failing intron positions is available at `http://genes.mit.edu/NielsenEtAl/`. These data constitute a genome-wide survey of high-confidence aligned intron positions and their patterns of conservation over at least 330 million years of evolution.

An example of an alignment of putative orthologs with three passing intron positions is shown in Figure 2-4A. In each passing intron position (black-edged rectangles), individual introns are labeled according to the classes previously outlined in Figure 2-1B. The first intron position is conserved across all four species, the second is a raw gain in N. crassa, and the third is present only in A. nidulans, and, because of the ambiguity in inferring gain or loss in this case, is classified as "Other". Examining the region around the one raw gained intron in N. crassa at the nucleotide level (Figure 2-4B) reveals a clean insertion of the intron sequence within a highly conserved region. The gained intron has consensus terminal dinucleotides `GT...AG` and a putative branch point sequence that matches the yeast consensus `TACTAAC` at six of seven positions. In addition, the fourth intron position of this set of orthologs was poorly aligned and excluded by our filters. All three passing positions (black-edged rectangles) display high amino acid sequence conservation on both sides flanking the intron, supporting the correctness of the alignment. In contrast, the failing intron position (unedged gray rectangle) is adjacent to a region of the alignment that lacks significant conservation. The $3'$ flank of this intron position displays considerable variation,

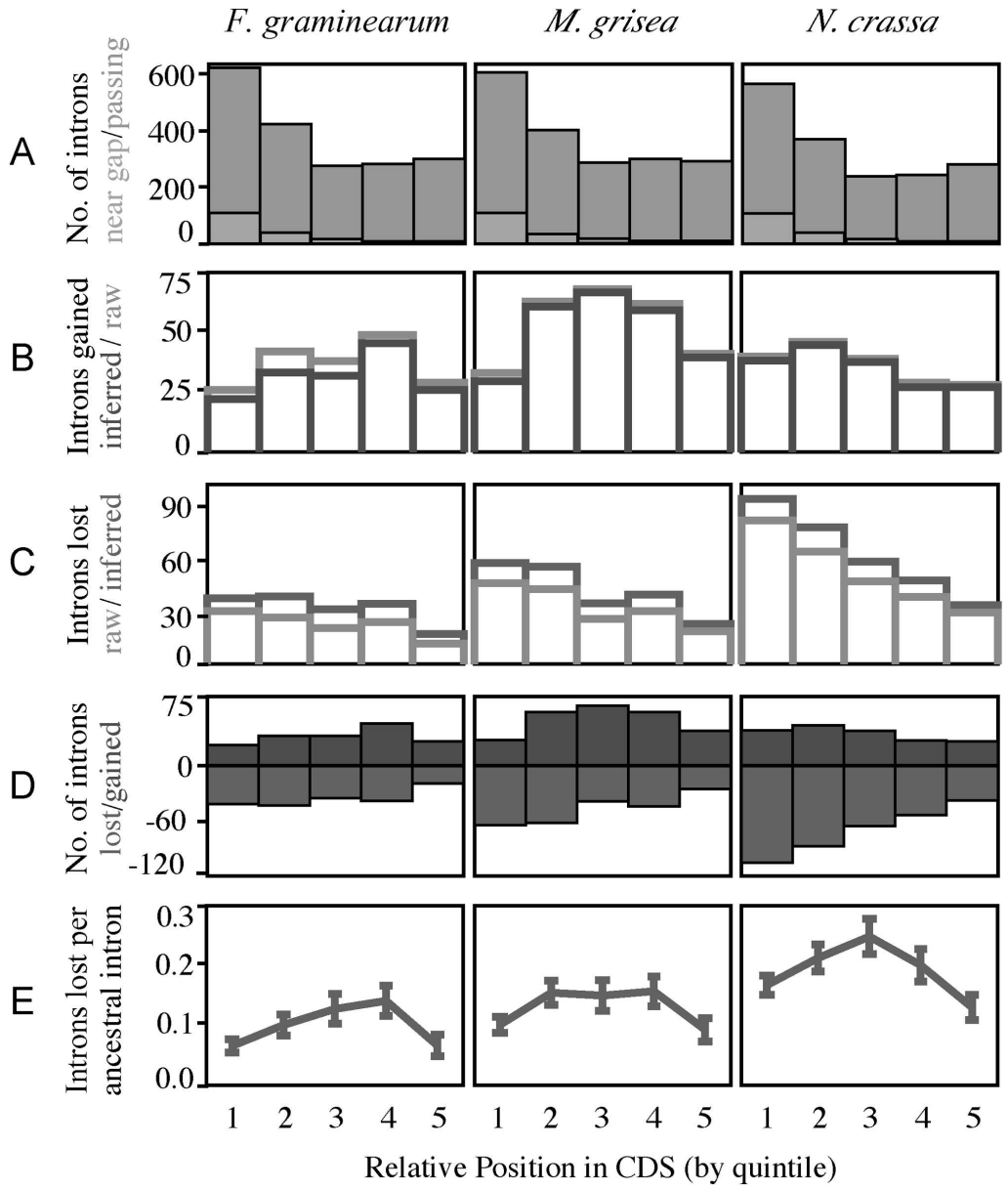


Figure 2-3: Positional Biases in Intron Gain and Loss. Relative intron positions were defined as the number of bases in the coding sequence upstream of the intron divided by the total length of the coding sequence. These relative positions were binned into five categories (quintiles), each representing one-fifth of the coding sequence length (quintiles numbered 1-5 on the x-axis). A, Introns passing quality filter (back) and introns adjacent to gaps in the protein alignment that were removed by our quality filter (front). B, Raw and inferred gains. Raw gains (back) are those introns present in exactly one organism (excluding the outgroup, *A. nidulans*). Inferred gains (front) are corrected for the estimated number of cases that arose by other combinations of gain and loss events. Inferred gains are thus slightly lower than raw gains. C, Raw and inferred losses. Raw losses (front) are those introns absent in the organism in question but present in at least one of its siblings (descendants of its parent in the phylogenetic tree) and one of its cousins (non-descendants of its parent). Inferred losses (back) are corrected for the estimated number of introns lost along multiple lineages, or gained and then lost. Inferred losses are thus slightly higher than raw losses. D, Number of introns gained (above 0) and lost (below 0) since last common ancestor. E, Intron loss rate at each position since last common ancestor (introns lost per ancestral intron). Error bars represent binomial standard deviation.

especially with respect to the M. grisea gene, which was predicted to have a much longer $3'$ coding region. In such an alignment region, it is difficult to distinguish true differences in intron conservation from potential annotation or alignment errors. Our filtering process thus eliminated this position from further analysis.

2.2.3 Calculation of Raw Gains and Losses

We calculated "raw gains" and "raw losses" by positional quintile for each organism other than the outgroup, *A. nidulans*. We defined raw gains as those introns present in only a single organism (see Figure 2-1B). We defined raw losses as those introns that are absent in the organism in question, present in some other descendant of the organism's parent (a "sibling"), and present in some non-descendant of the parent (a "cousin") (Figure 2-1B). Intron positions are considered conserved if present across all four organisms. Patterns of intron presence and absence that are not captured by the above definitions were excluded from the raw counts because of the ambiguity in inferring intron gain or loss events in such cases (marked as "Other" in Figure 2-1B).

2.2.4 Probabilistic Model of Intron Gain and Loss

Raw gain and loss counts are based on parsimony and may differ somewhat from the true number of gain and loss events. The set of raw gains may include introns that were lost in multiple lineages, thus over-counting the true number of gains in a given lineage. Similarly, the set of raw losses excludes introns lost in the given organism and also lost in all cousins or siblings (marked as "Other" in Figure 2-1B).

We used a probabilistic model to correct for these inaccuracies. Our model assumes that all loss and gain events occur independently and uniformly within each quintile. In particular, we assume Dollo's postulate [12]: any introns that align to the same position must have a common ancestor (no "double gains"), as in [36] and [48]. Our method differs from the Dollo parsimony method described in [16] and applied in [48] in that we do not artificially minimize loss events by assuming that gains occurred at the latest possible point in evolution. It also differs in that we allow

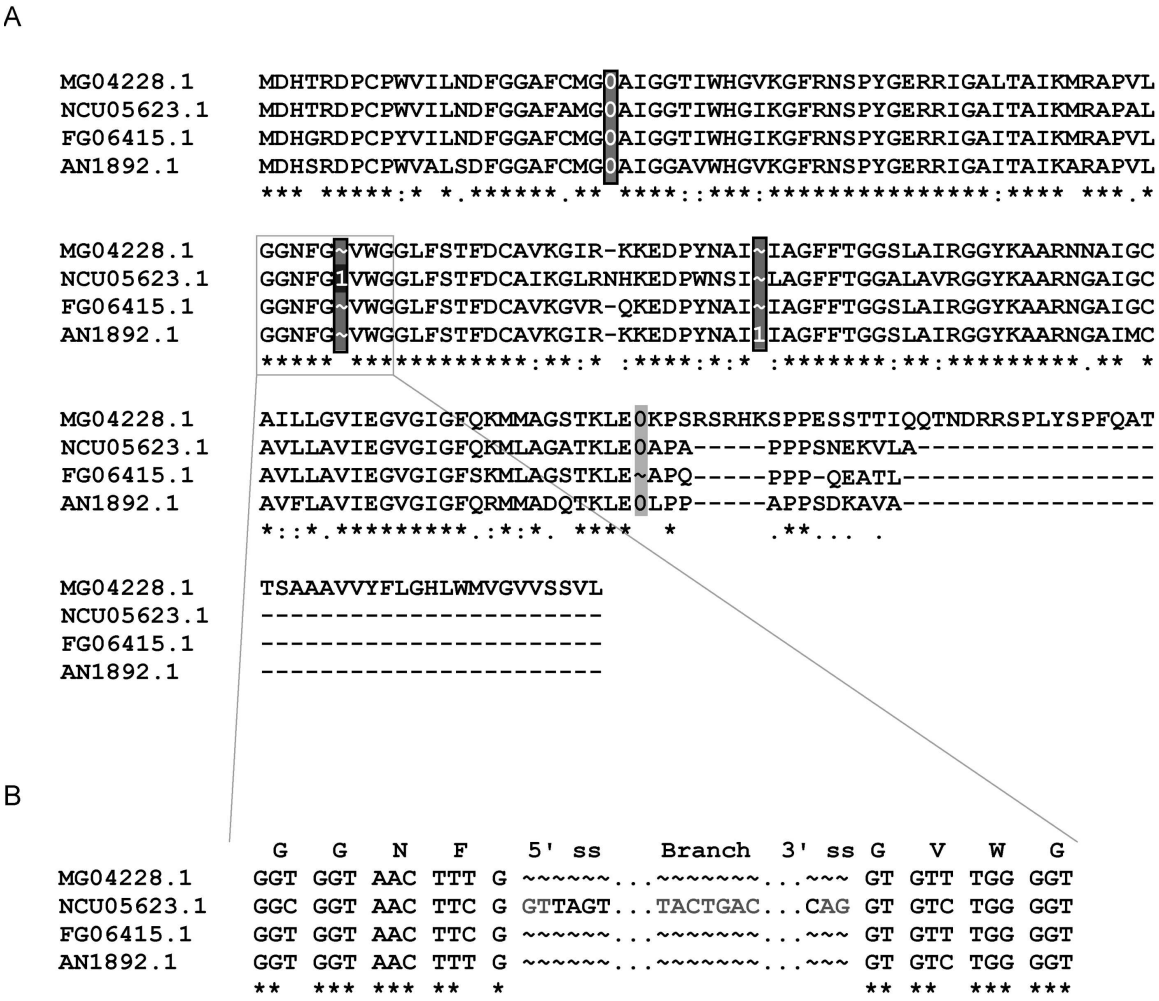


Figure 2-4: Example Ortholog Alignment. A, Alignment of protein sequences for orthologs MG04228, NCU05623, FG06415, and AN1892 with intron characters inserted. "0", "1", and "2" indicate the phase of an intron. A black-edged rectangle (first three intron positions) indicates an intron position passing our quality filters; the unedged rectangle (last intron) indicates an intron position that was removed by our filter. The first intron position was classified as "conserved" and the second as a "raw gain". The consensus (bottom) line characters are as follows: asterisk, identical residue in all four sequences; colon, similar residue; and period, neutral residue. B, Nucleotide alignment of the region flanking the gained intron in A. Putative 5′ and 3′ splice sites and a branch point sequence are shown from the intron.

different branches of the phylogenetic tree to have different rates of loss and gain. We applied our method separately to each of the five positional quintiles for each organism other than the outgroup, *A. nidulans*.

First we estimate two types of intron loss rates. The organismal loss rate, q, is calculated by dividing the number of raw losses in an organism by the total number of introns present in at least one sibling and at least one cousin. This represents the fraction of introns in the parent that did not survive to the present day in that organism. For instance, the organismal loss rate in *F. graminearum* is given by

$$q = \frac{AM + AN + AMN}{(AM + AN + AMN) + (AFM + AFN + AFMN)},$$

where AM, for example, represents the number of intron positions with an intron present in *A. nidulans* (A) and *M. grisea* (M) but absent from *F. graminearum* (F) and *N. crassa* (N).

The sibling loss rate, r is defined for a given organism as the fraction of introns in the parent that did not survive in any sibling. We define "sibling raw losses" for an organism as the number of introns that are present in the organism and at least one cousin but in no sibling. This quantity is then divided by the number of introns present in that organism and at least one cousin to give the sibling loss rate. For example, the sibling loss rate for *F. graminearum* is given by

$$r = \frac{AF}{AF + AFM + AFN + AFMN}.$$

We next correct the raw gains for each organism. Raw gains include some introns that were in fact lost in all but one lineage. We use the loss rates to calculate the expected number of these multiple losses, m, and subtract this quantity from the raw gains to obtain "inferred gains." To calculate m we first count B_0, the number of introns conserved in the organism and at least one sibling, but in no cousin. The quantities m and B_0 are related through the variable n_0, the number of introns present

in an organism's parent but not in any cousin, as follows:

$$m = n_0 r(1-q)$$
$$B_0 = n_0(1-r)(1-q). \tag{2.1}$$

This follows from our assumption of independent gains and losses. Thus, we can calculate the expected number of multiple losses as

$$m = \frac{B_0 r}{1-r}$$

We use the loss rates to estimate the number of introns in each organism's parent. To do so, we estimate separately the number of parental introns present in at least one cousin n_1, and the number not present in any cousin n_0 (introduced above). To estimate the size of the set of parental introns present in at least one cousin, we first count the subset of these introns that are presently observable. An intron is in this set if it is present in at least one cousin and at least one sibling, or is present in at least one cousin and in the organism in question. We call this number of introns B_1. By the assumption that gains and losses are independent, we have

$$B_1 = n_1(1-qr)$$

Using this relation and the one in equation 2.1 above, we calculate the number of introns in the phylogenetic parent as

$$n_{\text{total}} = n_1 + n_0 = \frac{B_1}{1-qr} + \frac{B_0}{(1-q)(1-r)}$$

Finally, we correct raw losses. Our definition of raw losses undercounts the true number by omitting those introns not conserved in at least one cousin and at least one sibling. Taking *F. graminearum* as an example, the true number of losses would also include some introns conserved in the patterns A, M, N, and MN. We calculate the number of inferred losses as $n_{\text{total}}q$.

This method can be extended to any phylogenetic tree and to any organism with at least one cousin.

2.2.5 Abundance of Intron Gains

One immediate conclusion stemming from our analysis is the importance of intron gain. A summary of all raw and inferred gains and losses is shown in Figure 2-3. Substantial numbers of gained introns were observed in all three organisms—more than 100 independent inferred gains in each lineage, with over 200 in *M. grisea* (Figure 2-3B). The total numbers of gains that have occurred in each genome are likely to be substantially higher, since only predicted orthologs in all four species were considered, and roughly a third of the introns in these genes passed our quality filters. Differences in intron dynamics between lineages are also apparent, with the numbers of gained and lost introns approximately balanced in *M. grisea* and *F. graminearum*, but with roughly twice as many losses as gains in *N. crassa* (Figure 2-3D). It is thus apparent from these data that the process of intron gain plays a significant role in intron evolution.

2.2.6 PRPP Synthetase Genes Display Lineage-Specific Increases in Intron Gain Rate

A striking example of intron gain occurs in a set of putative orthologous 1-phosphoribosyl-5-pyrophosphate (PRPP) synthetase genes. These genes encode a widely conserved protein that catalyzes the production of PRPP, a precursor in the nucleotide biosynthesis pathway. In contrast to the majority of orthologs that displayed fewer than two gained introns, the set of PRPP synthetase genes displayed a total of 22 raw gains (Figure 5A) that passed our alignment quality filters: six in N. crassa, 14 in M. grisea, and two in F. graminearum. The number of raw gains in the PRPP synthetase genes in *M. grisea* and *N. crassa* was significantly higher ($P < 3 \times 10^{-22}$ and $P < 4 \times 10^{-9}$ respectively) than the average for other genes analyzed, resulting in unusually large numbers of introns in these genes (Figure 2-5B). In comparison, the numbers of in-

trons in PRPP synthetase genes in available animal genomes were within the typical range for the respective organisms, e.g., five in *C. elegans*, and six in fruit-fly, human, mouse, rat, and *Fugu*. Thus the rate of intron gain for the PRPP synthetase gene in some fungi is unusually high. This gene represents an extreme example of the impact of intron gain and illustrates the variability of gain rates in different lineages.

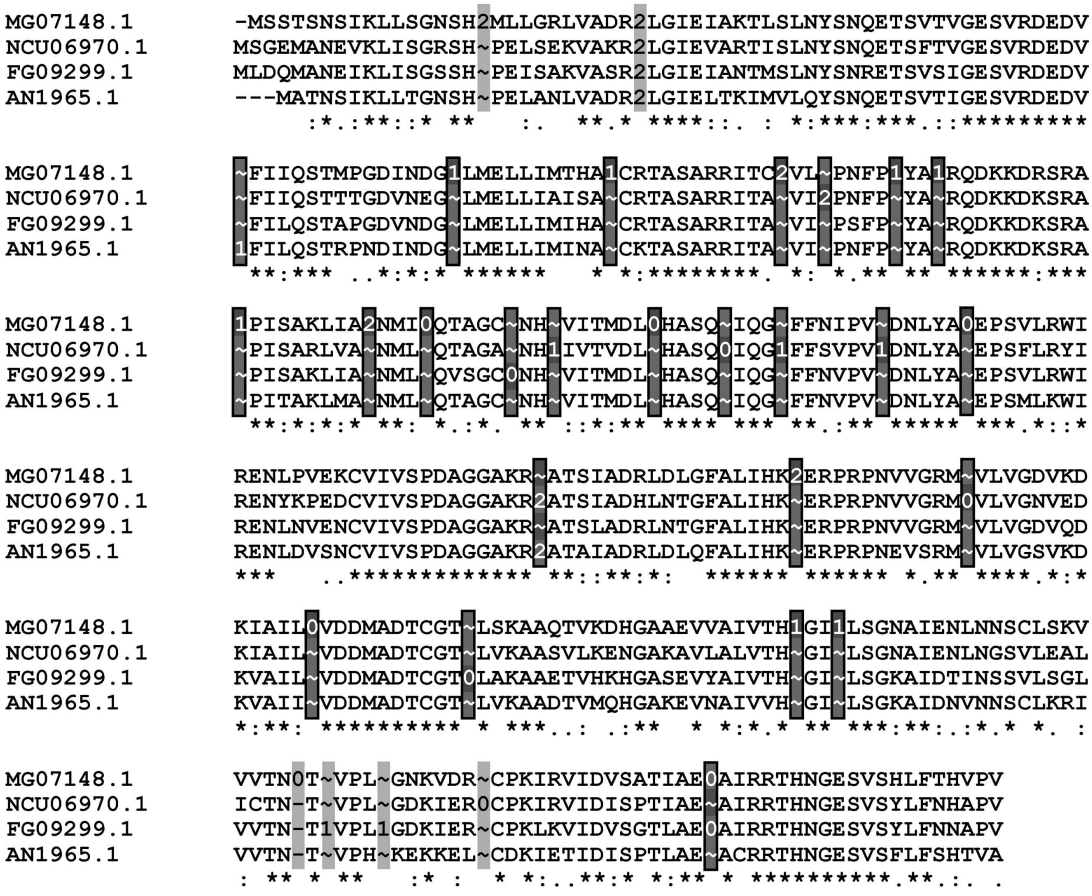
2.2.7 Absence of 3′ Bias in Intron Losses

To determine whether the pattern of intron loss in these fungi might account for the observed bias in intron position, we examined the pattern of loss as a function of position within the gene (see Figure 2-3E). Contrary to what would be expected if intron loss primarily involved homologous recombination of polyadenosine-primed reverse transcripts, the rate of intron loss tends to be lower, rather than higher, at the 3′ ends of genes. Moreover, the highest rates of intron loss occur in the middles of genes in all three organisms. We found no evidence that this pattern was affected by our filtering methods. These findings suggest either other mutational mechanisms (e.g., reverse transcription primed internally) or the presence of selective pressure to preferentially conserve introns near the 5′ and 3′ ends of genes.

2.3 Discussion

We developed a system that automatically identifies evolutionary and positional patterns of intron conservation on a genome-wide scale. The core of the system is a process for stringently filtering alignments of orthologous genes to exclude potential annotation or alignment errors. The result of the filtering process is a high-confidence set of aligned intron positions. Differences in intron conservation at each individual position can be characterized as gains or losses (or ambiguous) based on parsimony. However, this does not accurately account for the possibility of multiple gain or loss events. We have developed a probabilistic model that allows for multiple events, providing a corrected estimate of the total number of gains and losses within the dataset. Our probabilistic method allows for a more accurate assessment of rates of gain and

A



B

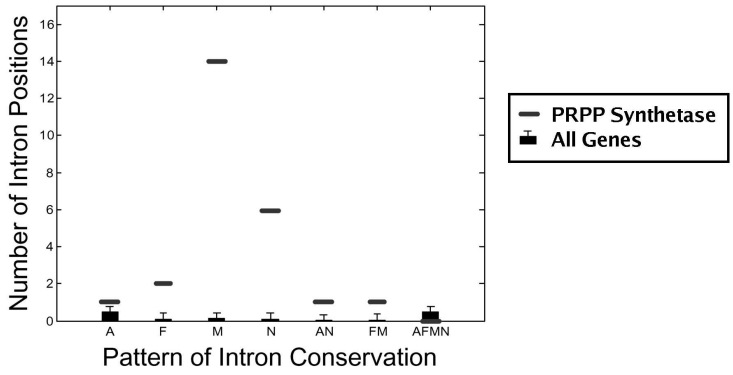


Figure 2-5: Intron Conservation in the PRPP Synthetase Gene. A, Alignment of PRPP synthetase putative orthologs MG07148, NCU06970, FG09299, and AN1965. A black-edged rectangle indicates an intron position passing our quality filters, whereas an unedged gray rectangle indicates an intron position that was removed by our filter. Blue boxes mark raw intron gains, red boxes indicate raw intron losses, and gray boxes within black-edged rectangles highlight all other introns. We manually corrected an annotation error in the first intron of the last row of the alignment. B, Phylogenetic conservation pattern of introns in the PRPP synthetase gene. Each passing intron position was categorized as being present in *A. nidulans* (A), *F. graminearum* (F), *M. grisea* (M), *N. crassa* (N), *A. nidulans* and *N. crassa* (AN), *F. graminearum* and *M. grisea* (FM), or all four organisms (AFMN). There are no passing cases of conservation in three or four species. The number of introns in each category is shown with a purple line. The black error bar plot shows the mean and standard deviation for each category for all 2,008 ortholog sets after fitting to a Poisson distribution (see Section 2.4.5). The number of introns in *M. grisea* and *N. crassa* is significantly higher, at the $P < 1 \times 10^{-9}$ level.

loss. In our dataset, allowing for multiple events results in only modest corrections to the rates estimated using parsimony.

Our analysis demonstrates a significant role for intron gain over the past few hundred million years in the fungi analyzed. Previous analyses of specific gene families have provided evidence of specific instances of gained introns [33], [47], [22], [43]. However, the relative importance of intron gain versus loss is not well understood. Recent large-scale analyses have suggested that intron gain may play a predominant role in shaping gene structures [43], although lineage-specific differences are apparent [48]. In particular, intron gain appears to occur rarely if at all in mammalian genes [49]. Our data suggest that intron gain is a significant driving force in the evolution of genes in fungi. In *F. graminearum* and *M. grisea* the number of introns gained was on par with the number lost and similar in magnitude to the number of introns gained in *N. crassa*.

The mechanisms underlying intron gain are not known. We analyzed the set of predicted intron gains for possible signatures that might shed light on this process. No statistically significant bias was detected in the positions of gained introns along the coding sequence (see Figure 2-3 and data not shown). Similarly, no preferred insertion site sequence was detectable and no significant phase bias for gained introns was observed (data not shown). The lack of an insertion site preference and absence of significant phase bias for gained introns in fungi is consistent with previous investigations and may set fungi apart from other organisms [43].

Our data further indicate that intron gain can vary substantially between different gene families in a lineage-specific fashion. The PRPP synthetase gene is a particularly striking example, exhibiting significant increases in gained introns in two of the four lineages investigated. Moreover, the paucity of intron positions shared between *N. crassa* and *M. grisea* suggests the possibility of independent increases in gain rate in the two species. Alternatively, the apparent high intron gain rate exhibited by this gene may have arisen just prior to the last common ancestor of *N. crassa* and *M. grisea*. Although it is premature to speculate about possible mechanisms, one possibility is that a factor or factors responsible for intron insertion evolved to associate

with the PRPP synthetase gene locus, transcript, or message at this point, leading to a higher rate of intron insertion in this gene.

Finally, our results do not support the mechanism commonly proposed to account for the 5′ positional bias of introns in intron-poor organisms [35]. Contrary to what would be expected if intron loss primarily involved recombination of polyadenosine-primed reverse transcripts, the rate of intron loss tends to be lower at the 3′ ends of genes. Instead, the highest rates of intron loss occur in the middles of genes in all three organisms. (This result is consistent with the results of [49], an analysis of intron evolution in mammals. Although the report describes only six instances of loss, in each case it was an internal intron.) The preference for internal introns may reflect a process of reverse transcription primed internally. Alternatively, there may be pressure to preferentially conserve introns near the 5′ and 3′ ends of genes. In particular, there is strong evidence for a functional role for the 5′-most intron in many genes. What remains clear is that the pattern of loss in these fungi over the last 330 million years cannot be explained solely by a mechanism involving 3′-end-primed reverse transcription of spliced messages. Instead, fungal intron dynamics appear to reflect a more complex interplay between intron gain and loss, an interplay that is likely to shape intron evolution in other eukaryotes.

2.4 Materials and Methods

2.4.1 Sequences and Annotations

All sequences and annotations were taken from the Broad Institute Fungal Genome Initiative website (`http://www.broad.mit.edu/annotation/fungi/fgi`). The following datasets were used: *A. nidulans* (Assembly 1, 18 February 2003), *N. crassa* (Assembly 3, 1 February 2001), *F. graminearum* (Assembly 1, 11 March 2003), and *M. grisea* (Assembly 2, 18 July 2002).

2.4.2 Ortholog Identification

A group of four proteins, one from each organism, was considered an ortholog set if each pair was a pairwise best bidirectional BLAST hit in the respective genomes, and all the BLAST hits overlapped by at least 60% of the length of the longest protein. This yielded 2,073 sets of orthologs (out of an average of 10,500 genes in the four organisms). We repeated our analysis, requiring that each best bidirectional hit also be the only BLAST hit in each genome (spanning 60% the length of the longest protein). This protocol yielded only 1,178 ortholog sets, but gave qualitatively similar results for intron gains and losses (data not shown).

2.4.3 Ortholog Alignment

The proteins in each ortholog set were aligned using ClustalW 1.82 [7], and intron position characters were inserted into the alignments, using "0", "1" or "2" to indicate the intron phase. Phase 0 intron characters were inserted between the amino acids coded for by the codons adjacent to that intron, and phase 1 and 2 intron characters were inserted immediately following the amino acid coded for by the codon interrupted by the intron. If an intron was not present in all the sequences at a given position, special intron gap characters "~", were inserted in the other sequences in order to maintain the downstream amino acid alignment. A total of 9,352 intron positions were aligned. At only 28 (0.3%) of these positions were introns of different phases aligned, making it reasonable to ignore "phase shifting" in our analysis.

2.4.4 Alignment Filtering

Regions of low alignment quality were eliminated with a filter that required at least 30% identity and 50% similarity in a window of ten residues on each side of the intron position. These parameters were determined following manual classification of a set of 181 randomly selected intron positions as "clearly homologous," "ambiguous/possibly homologous," or "non-homologous" (see Figure 2-2B). Using the parameters above, 92% of the homologous positions, 29% of the ambiguous positions and only 2% of the

non-homologous positions passed the filter.

To further exclude likely annotation and alignment errors, intron positions were also filtered by eliminating positions adjacent to gaps in the amino acid alignment and by eliminating positions with nearby introns but low evidence of homology in the intervening sequence. It is possible that some of these positions may in fact reflect intron gain or loss events that occurred simultaneously with coding sequence insertion or deletion. However, removing these positions did not significantly impact our results, as the number of positions adjacent to gaps was only about one-tenth of the number of positions that passed the quality filter, and the introns removed did not have an apparent positional bias (see Figure 2-3A).

2.4.5 Statistical Significance of High Gain Rate in PRPP Synthetase

We modeled the number of gains in a particular organism as a Poisson distribution under two different null hypotheses. One null hypothesis was that the gains were spread uniformly across all genes. The other was that the number of gains in each gene was proportional to the length of the gene. In the first case the Poisson parameter λ is given by the total number of raw gains observed in that organism divided by the total number of ortholog sets ($P < 3 \times 10^{-22}$ for *M. grisea*, $P < 4 \times 10^{-9}$ for *N. crassa*, and $P < 0.007$ for *F. graminearum*). In the second case λ is given by the total number of raw gains observed in that organism multiplied by the length of that gene in amino acids and divided by the total number of amino acids in all genes in that organism ($P < 7 \times 10^{-25}$ for *M. grisea*, $P < 3 \times 10^{-10}$ for *N. crassa*, and $P < 0.003$ for *F. graminearum*). We reported the less significant of the two P-values in the results.

2.5 Acknowledgments

This work was completed in collaboration with Cydney Nielsen, then at the Broad Institute. She created the ortholog sets and sequence alignments with inserted intron

characters. She also wrote a library of PERL methods to handle these data structure, and made early versions of the filtering protocol to which I made significant changes. My contributions included the final version and implementation of the filtering protocol, the mathematical modeling, the discovery of PRPP synthetase, and the statistics and data visualization.

This chapter is nearly identical in content to our paper [38].

Chapter 3

Discovery of Interacting Regulatory Elements Using Collocation

Abstract

The sequence determinants of the RNA splicing reaction in human include canonical elements at the beginning and end of every intron, but these are not sufficient to specify the splice sites of most exons. Other elements, known as *splicing regulatory elements*, can function within exons or flanking introns to modulate inclusion levels and exon boundaries. A number of computational and experimental screens have been successful in defining a library of known regulatory elements. However, there are many known interactions between sequence elements that might not function alone, and therefore would have been missed by these screens. We present here a method to detect interacting motifs in pairs of sequences and apply it to search for motif pairs functioning between the beginning and end of introns. The method controls for the known G+C heterogeneity of the human genome, which otherwise can give false signals. The method is successful in recapitulating the U12 splice sites, elements marking the beginning and end of the 0.2% of human introns spliced by the minor spliceosome,

and also identifies a novel motif pair. This motif pair collocates across constitutive introns as well as distal splice sites of introns flanking alternatively included exons, and also is enriched in very long introns, suggesting that it functions to suppress the inclusion of intervening exons. Experiments with a mini-gene construct confirm this hypothesis.

3.1 Introduction

The sequence determinants of splicing are still not fully understood. While the sequences occurring at the splice sites have been well characterized (see Figure 1-6 and [65]), there are many examples of these sequences appearing together and yet not being included in any known transcript, and there are also some exons with weak splice sites that are still constitutively included. One of the reasons that splice sites alone cannot predict exon/intron structure is that other sequence elements are important in determining splice choices.

To date, a large number of these elements have been found and characterized. The most well-studied classes are the *exonic splicing enhancers* and *silencers* (ESEs and ESSs). These are sequences that appear within exons and enhance (or suppress) their inclusion into the final transcript. Similarly, *intronic splicing enhancers* and *silencers* appear in introns flanking exons and either enhance or suppress the exons' inclusion. Collectively, sequences in these four classes are known as *splicing regulatory elements* (SREs).

Various *in vitro*, *in vivo* and *in silico* approaches to high-throughput identification of such elements have been carried out. Experimental screens have used either screening of random oligonucleotide libraries in cell-based assays [61, 11] or SELEX [59] to determine sequences that function to control splicing. The assays for this activity have been both functional (inclusion or skipping of a reporter exon) and binding (to a known trans-factor).

The computational approaches have been based on k-mer statistics. One approach to identifying ESEs, described in section 1.6.2, looked for hexamers enriched in weak

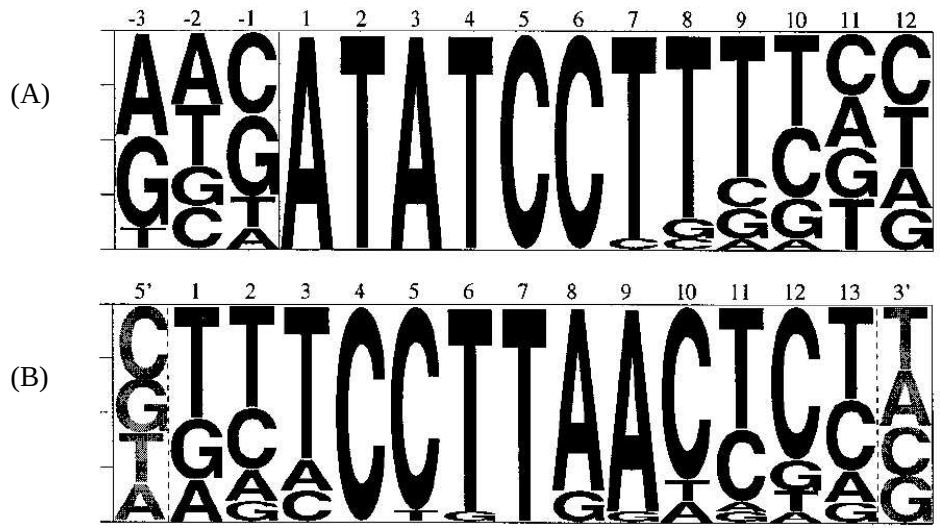


Figure 3-1: The U12-dependent 5′ splice site (A) and branch site (B). The first three nucleotides from (A) are from the upstream exon and the rest is from the beginning of the intron. At the time of publication it was not known that some u12 introns begin with `G` instead of `A`. Figure modified from [5].

exons relative to strong exons and in exons overall relative to introns [15]. Another paper looked instead for octomers enriched in internal non-coding exons relative to both pseudoexons and intronless genes [66].

All of these methods aimed to discover *isolated* elements. If an element only functioned in concert with a second element then it might not be uncovered. One such example would be the U12 5′ splice site and branch site (Figure 3-1). These sequences are found respectively at the beginning and end of introns spliced by the so-called minor spliceosome. This is a second spliceosome that shares some molecular components with the major spliceosome, but for which some key components are different. In particular, the specificity of the 5′ and 3′ splice sites are very different, and in general a U12 splice site and a U2 splice site are incompatible within the same intron. As such, either one of the U12 splice sites is functionless alone, but together they can promote the excision of an entire intron.

Another example of a pair of elements that function together are the binding sites of the polypyrimidine tract binding protein (PTB). These sites are known to appear flanking some alternatively skipped exons. PTB is a multi-domain protein in which each domain has a similar binding specificity. When two domains bind to `C`/`U` -rich

elements flanking an exon, PTB can cause a "looping out" of the exon, effectively preventing its inclusion in the mature transcript [40].

Knowing these as well as other examples, we set out to identify interacting SREs *in silico*. Our philosophy was that such motifs should "collocate", that is, they should co-occur more often than expected by chance. We developed a statistical framework that can identify collocating motifs while controlling for artifacts unrelated to splicing: the G+C heterogeneity of the human genome, repetitive elements, and paralogous genes. We showed that, in a variety of control sets, similarly enriched pairs of motifs could not be found. Finally, we explored computationally a novel detected pair and confirmed it effect on splicing in a mini-gene reporter transfected into HeLa cells.

The statistical method we develop is general enough to be applied to a number of different data sets, for example to identifying motif pairs that flank polyadenylation or branch point sites, or that co-appear in promoter regions and their distant enhancers.

3.2 Results

3.2.1 Simple Collocation Exhibits G+C Bias

Our method begins with a set of N pairs of sequences. These sequences typically bear some biological relation to each other: beginning and end of introns, intron and exon flanking $5'$ splice sites, etc. We choose some oligonucleotide length k to study. For each pair of k-mers x, y we would like to test if x and y tend to collocate. Let $C(x, y)$ denote the number of x-y *collocations*, sequence pairs that contain x in the first sequence and y in the second.

Since $C(x, y)$ is affected by the abundance of x in the first sequence set and y in the second, we take this into account in deciding if x and y significantly collocate. In particular, let $n_1(x)$ denote the number of sequences containing x in the first set, and similarly for $n_2(y)$. Next define the frequencies of the hexamers by normalizing their counts, $f_1(x) := n_1(x)/N$ and $f_2(y) := n_2(y)/N$. Then under the null hypothesis that occurrences of x and y in their respective sets are independently distributed, the

expected number of collocations is

$$\mathbb{E}C(x,y) = Nf_1(x)f_2(y). \tag{3.1}$$

To test if $C(x,y)$ is significantly higher than $\mathbb{E}C(x,y)$ we model the collocations as a Poisson process, so that $C(x,y)$ is distributed as a Poisson random variable with parameter $\lambda = \mathbb{E}C(x,y)$. This approximation is very good so long as the frequency of a single collocation under the null hypothesis $f_{\text{coll}} = f_1(x)f_2(y)$ is not too big. Averaged over all hexamers, $f_i(x)$ is about $4^{-k}L$ where L is the length of the sequences in the sequence set, so f_{coll} averages to about $L^2 4^{-2k}$. In this project we work with $k \geq 4$ and $L \leq 80$, so this quantity is at most 1%. Thus we can calculate a P value for significance of collocation using

$$\begin{aligned} P &= P(X \geq C(x,y)) \\ &= 1 - P(X < C(x,y)) \\ &= 1 - P(X \leq C(x,y) - 1) \\ &= 1 - F_{\text{pois}}(C(x,y) - 1, \mathbb{E}C(x,y)), \end{aligned} \tag{3.2}$$

where X is a Poisson random variable with parameter $\mathbb{E}C(x,y)$ and $F_{\text{pois}}(\cdot,\lambda)$ denotes the cumulative distribution function for the Poisson distribution with parameter λ.

Using this method we calculated the significance of collocation between the first and last 80 nucleotides of introns (see section 3.4) for each of 4^{12} pairs of hexamers. At a significance cutoff of $P = 4^{-12}$ we would expect a single false positive. Indeed, we found 72366 significant collocations between 1377 hexamers at the 5′ splice site and 1374 hexamers at the 3′ splice site.

Figure 3-2 depicts these hexamers as nodes, colored by their G+C content, and the collocations between them as edges. From this plot we can see an obvious partition of the hexamers into a G+C-rich (light) and an A+T-rich (dark) set. This has been interpreted to mean that there are two spliceosomes, one corresponding to each set, but we believe instead that this is an artifact due to the G+C heterogeneity of the

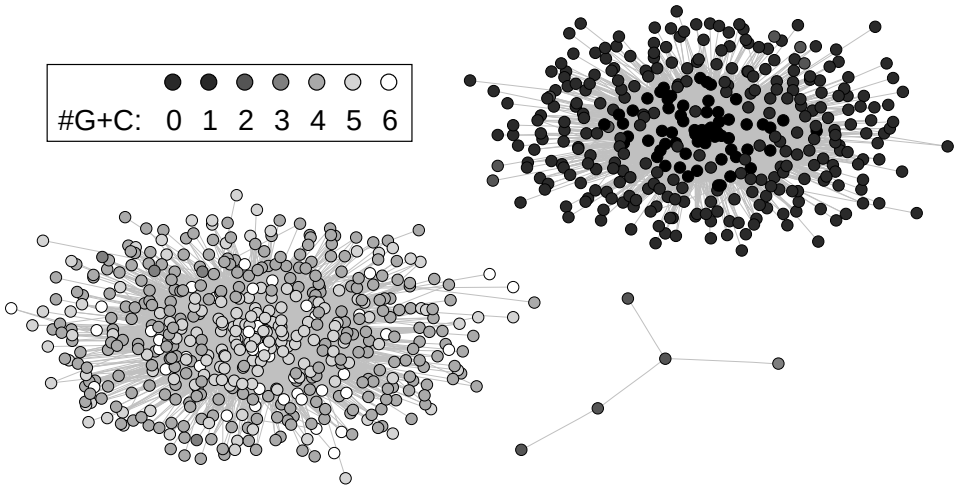


Figure 3-2: Significant collocations between the beginning and end of introns under a null hypothesis of independence. Each node represents a hexamer, colored according to its C+G content, and each edge represents a significant collocation at $P \leq 4^{-12}/100000$. We choose such a conservative P value cutoff, at which we expect to see no false positives, for visualization purposes only. Note that this plot does not show the sets (beginning or end of intron) to which the hexamers belong. At this P value cutoff 848 nodes (21% of all hexamers) are joined by 21,389 collocations.

human genome.

3.2.2 G+C Heterogeneity Causes Many Collocations

It has been known for several decades that many genomes, especially those of warm-blooded mammals, exhibit nonuniform sequence composition. The biggest variation is in G+C content. Indeed, if one performs principle component analysis (data not shown) on the dinucleotide content of samples of the human genome, the first principle component is composed of approximately equal amounts of the four dinucleotides made of just G and C, (CC, CG, GC and GG) and approximately equal but opposite amounts of the four made of just A and T (AA, AT, TA and TT). The other eight dinucleotides contribute very little to this component, supporting the notion that the most important sequence heterogeneity in the human genome is G+C content.

Why would such heterogeneity lead to collocations? Because of the way it is organized. G+C content is tightly correlated across stretches of the genome extending much beyond the length of a single gene. The heterogeneity occurs between different

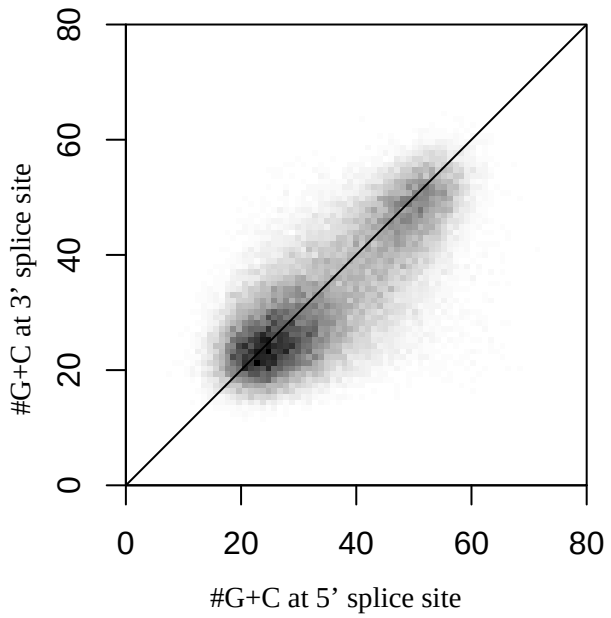


Figure 3-3: G+C content in the first 80 nucleotides (x-axis) and last 80 nucleotides (y-axis) of introns is correlated. Darker squares correspond to more introns (53326 total). The diagonal line is a plot of $y = x$ for reference.

chromosomes or different parts of the same chromosome. One way to see this correlation is to plot a scattergram of the G+C contents of the sequence pairs in our intron begin/end set. Figure 3-3 shows this as a heat map.

How might G+C heterogeneity have influenced the collocation results shown in Figure 3-2? In fact it could explain nearly all of them, because G+C-rich hexamers are more frequent at both ends of introns in G+C-rich regions of the genome, and A+T-rich hexamers are more frequent in A+T-rich regions of the genome, even though this is unrelated to splicing. Statistically speaking, the assumption of independence between in the two sets is invalid.

To examine this possibility, we *co-GC shuffled* the sequence pairs. To each sequence pair we associate the pair (s_1, s_2) of G+C contents, and we then reassign the pairings in such a way as to preserve the total number of pairs with each co-GC content (see Figure 3-4). In our example, s_1 represents the number of G+C in the first 80 nucleotides of the intron and s_2 represents the number in the last 80 nucleotides. This type of shuffling controls for the sequences in the sequence sets as well as the correlated bias in G+C content exhibited by the sequence pairs.

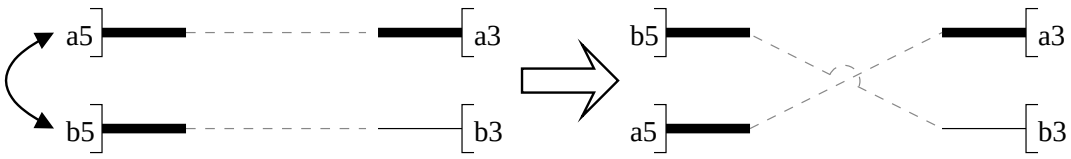


Figure 3-4: co-GC shuffling. (Left) Two hypothetical introns, A and B, with 5′ splice site (a5/b5) and 3′ splice site (a3/b3). Intron A is G+C-rich near both splice sites (thick lines) and intron B is G+C-rich near 5′ splice site, but G+C-poor near 3′ splice site (thin solid line). Since both introns are G+C-rich near their 5′ splice sites we can swap those splice sites (double arrow). (Right) co-GCshuffled introns. The beginning of intron B (b5) is now paired with the end of intron A (a3), top, and the beginning of intron A (a5) is now paired with the end of intron B (b3), bottom. Overall co-GC content is preserved. See section 3.4.3 for more details.

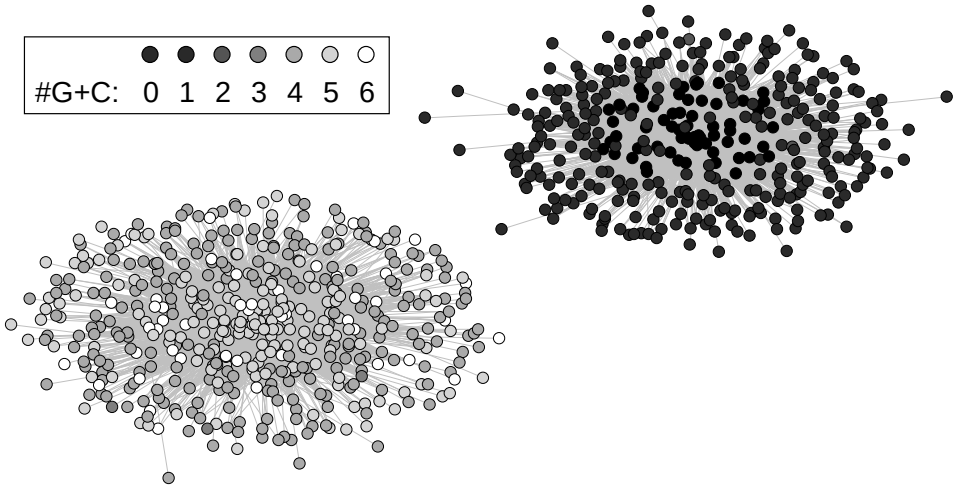


Figure 3-5: Significant collocations between co-GC shuffled intron beginnings and ends. Same as Figure 3-2 except for shuffling. At the same cutoff of $P \leq 4^{-12}/1000000$ we still have 840 hexamers (unshuffled: 848) and 18,645 collocations (unshuffled: 21,389).

Indeed, we see that even in this co-GC shuffled set, we still find 67,068 significant collocations between 1351 hexamers at the 5′ splice site and 1312 hexamers at the 3′ splice site (Figure 3-5). Therefore over 90% of the signal from the real data can be explained simply by G+C heterogeneity, and therefore is not likely to be related to splicing.

3.2.3 Stratification Controls for G+C Heterogeneity

We next set out to find a method that would control for the sequence heterogeneity that was obscuring any signal that might be present. We knew that for a restricted

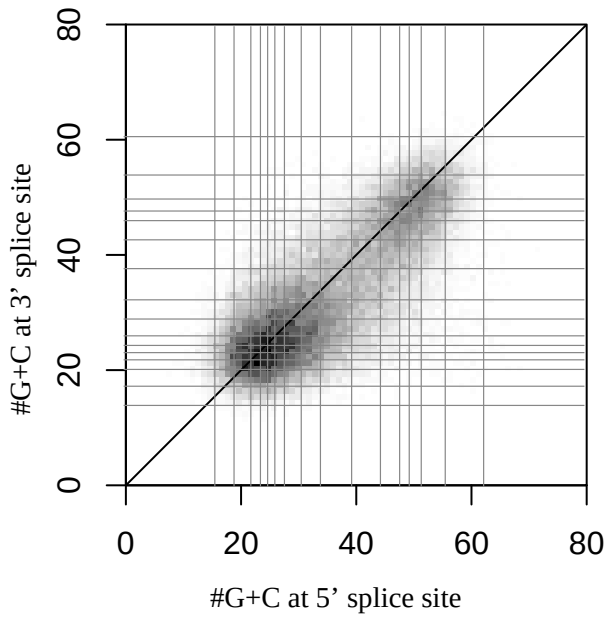


Figure 3-6: A hypothetical co-GC binning of the intron begin/end sequence pairs. Gray lines correspond to bin boundaries overlaid on Figure 3-3. Note that the boundaries are closer together around 23 and 50 G+C where most introns fall. The binning method chooses boundaries so that each one-dimensional bin has the same number of sequence pairs in it. The co-GC bins would correspond to the cells in the grid laid out by the gray lines, and show great variation in number of sequence pairs.

set of sequence pairs that all had similar G+C content the assumption of independence of the pairs was not unreasonable. Therefore, we doubly binned the sequence pairs according to their co-GC content (s_1, s_2), where s_1 denotes the G+C content of the first sequence and s_2 denotes the G+C content of the second sequence. We created 400 bins, each with a double index (b_1, b_2) $(1 \leq b_i \leq 20)$ corresponding to binnings of G+C contents of the first and second sequences, respectively. Thus this *co-GC binning* yielded 400 bins, each one containing sequence pairs whose first sequences had similar G+C contents and whose second sequences had similar G+C contents (Figure 3-6).

For each bin (b_1, b_2) we can imagine analyzing the data in a similar way as we did for the entire set. Let N^{b_1,b_2} denote the number of sequence pairs in the bin (b_1, b_2), and let $C^{b_1,b_2}(x, y)$ denote the number of sequence pairs in that bin containing k-mers x and y respectively. We define $n_1^{b_1}(x)$, the number of first sequences in G+C bin b_1 containing x, and similarly $n_2^{b_2}(y)$, the number of second sequences in G+C bin b_2 containing y. Letting $N_1^{b_1}$ denote the total number of first sequences of G+C content b_1 and $N_2^{b_2}$ the number of second sequences of G+C content b_2, we can define the

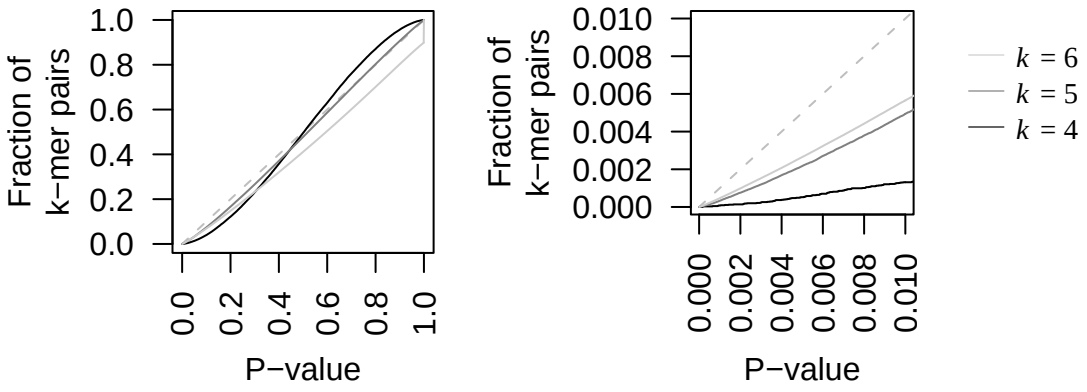


Figure 3-7: Probability plots of co-GC shuffled controls. x-axis, significance level. y-axis, fraction of data significant at that level. Left, entire range of P-values. Right, zoom in on smallest P-values. $k = 4, 5$ and 6 are colored in black, dark gray, and light gray respectively.

k-mer frequencies by $f_1^{b_1}(x) := n_1^{b_1}(x)/N_1^{b_1}$ and $f_2^{b_2}(x) := n_2^{b_2}(x)/N_2^{b_2}$. Putting all this together gives the analog of equation 3.1,

$$\mathbb{E}C^{b_1,b_2}(x,y) = N^{b_1,b_2} f_1^{b_1}(x) f_2^{b_2}(y). \tag{3.3}$$

We then use the property of the Poisson distribution that adding independent Poisson random variables corresponds to adding their parameters—if $X_i \sim Poi(\lambda_i)$ then $\sum_i X_i \sim Poi(\sum_i \lambda_i)$. Since the total number of x-y collocations is such a sum, $C(x,y) = \sum_{b_1,b_2} C^{b_1,b_2}(x,y)$, it should take a Poisson distribution with parameter $\mathbb{E}C(x,y) = \sum_{b_1,b_2} \mathbb{E}C^{b_1,b_2}(x,y)$. Calculating this sum of expectations and putting it into formula 3.2 then gives us the significance of collocation of the k-mers x and y.

Does this method in fact control for the G+C heterogeneity of the human genome? We can go back to our co-GC shuffled controls from Figure 3-5. In this case, there are no significant collocations at $P \leq 4^{-2k}$ for $k = 4, 5$ or 6. Another way to see that the new method accurately calculates P values is to look at a probability plot, which shows how the fraction of significant collocations (out of all possible 4^{2k} collocations) depends on the P value cutoff. For the shuffled data, this plot should be roughly a straight line, indicating that we get the expected number of false positives (Figure 3-7, left). Zooming in on the small P values (for example up to 1%) shows that in fact the method is somewhat conservative (Figure 3-7, right). In this regime the controls

show even fewer significant collocations than expected by chance. This is because a perfectly straight line in a probability plot only follows under the assumption that P values are uniformly distributed under the null hypothesis, and this only holds for continuous distributions. In short, there is no detectable signal in these shuffled controls—we have effectively controlled for G+C heterogeneity.

3.2.4 GC-Stratification Reveals Three Groups of Collocations Between Beginning and End of Introns

We have shown that our negative controls now show no significant collocations. What about our real data? Figure 3-8 displays the significant GC-stratified collocations between the beginning (5′SS) and end (3′SS) of introns. The three columns correspond to $k = 4, 5$ and 6 respectively. Note that the collocations shown here are at a P value that would give a single expected false positive, whereas the cutoff in previous figures (unstratified) was $1/1{,}000{,}000^{\text{th}}$ of that, and we still see the signal decrease by several orders of magnitude. Again, the data clusters very cleanly, with no edges between clusters. Three different groups of motifs are shown in the figure, indicated by different line styles.

The first cluster corresponds to splice sites of the U12-type introns (dashed line, $k = 6$ only). The set near the 5′ splice site matches the consensus for 5′ splice sites of U12-type introns perfectly if one allows G at position +1 of the intron, which is present in about 2/3 of U12-type introns. The set near the 3′ splice site matches the consensus for branch sites of U12-type introns (Figure 3-1). This cluster has 11 collocations, and the non-stratified method also identifies 11 U12 collocations at the same P value, 10 of which are shared between the two methods (data not shown). Coincidentally, the small 5-node cluster in the bottom right of Figure 3-2 (significant collocations calculated by the non-stratified method at the more stringent $P \leq 4^{-12}/100000$ cutoff) is formed of members of this motif. On the other hand, the total number of significant collocations decreased roughly 1000 times upon GC-stratification. Thus GC-stratification purifies true collocations from false positives.

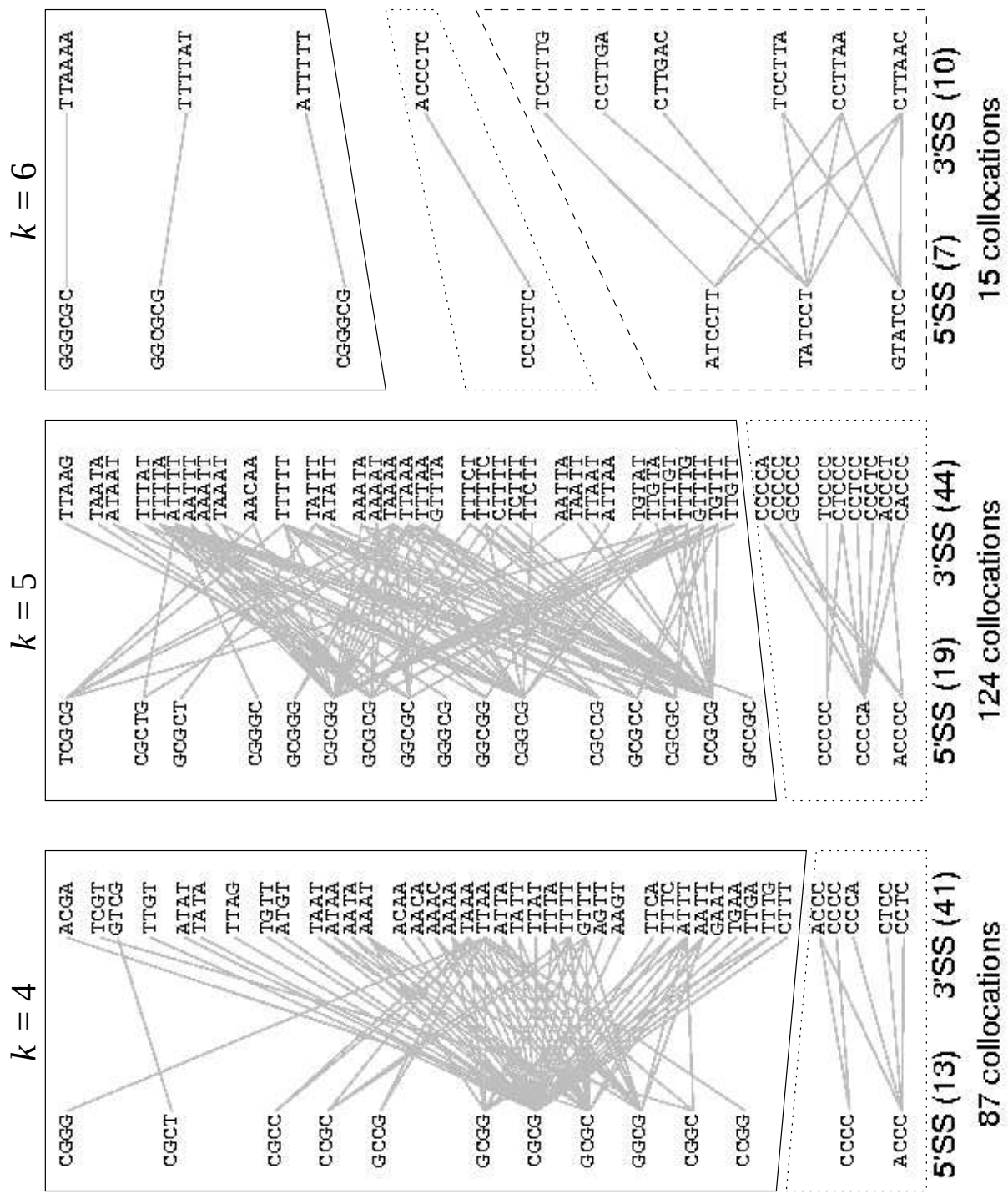


Figure 3-8: Significant GC-stratified collocations between the beginning and end of constitutive introns. Three columns correspond to $k = 4, 5, 6$ with P value cutoffs $P \leq 4^{-k}$, corresponding to a single expected false positive. Only k-mers participating in significant collocations are shown. Total number of k-mers at beginning (5′SS) and end (3′SS) of intron shown, as well as total number of collocations. Dashed line, U12 cluster. Dotted line, C-rich cluster. Solid line, GC/AU cluster.

An interesting aspect of this cluster is that it only appears at $k = 6$. A likely explanation for this is that the true motif is a long one and U12-type introns are very rare (about 0.2% of all introns). Thus the signal for collocation may be diluted by non-functional short submotifs for $k < 6$.

The second cluster is the C-rich cluster (dotted line). These also appear in the non-stratified results. We have not undertaken a more thorough study of their possible function, however Crich elements have been previously identified as putative intronic splicing enhancers [65].

We called the final cluster the GC/AU cluster (solid line). At the beginning of introns it features a G+C-rich motif with no other apparent structure, and at the end it features an A+U-rich motif, which seem to have some preference for stretches of A or U. This motif is present for all three values of k. Unlike the other two motifs, it is not found at all by the non-stratified method.

3.2.5 Control Sequence Sets Show No Collocations

Since the GC/AU cluster is not significant in the non-stratified method, one might ask whether it is an artifact of stratification. In other words, since G+C-rich and A+U-rich intron ends rarely co-occur, the expected number of collocations for G+C-rich and A+U-rich k-mers is lower under GC-stratification than non-stratification. This would be a side effect of stratification, since the original aim was to raise the expected number of collocations between G+C-rich and G+C-rich or A+U-rich and A+U-rich k-mers.

Three observations argue against this interpretation:

- The co-GC shuffles (Figure 3-7) do not show the GC/AU cluster as significantly collocating.
- The flipped motifs, with the G+C-rich motif at the $3'$ splice site and the A+U-rich motif at the $5'$ splice site, is not significant.
- There seems to be a tendency towards repeated A or repeated U in the A+U-rich motif, rather than a general significant collocation of all A+U-rich motifs.

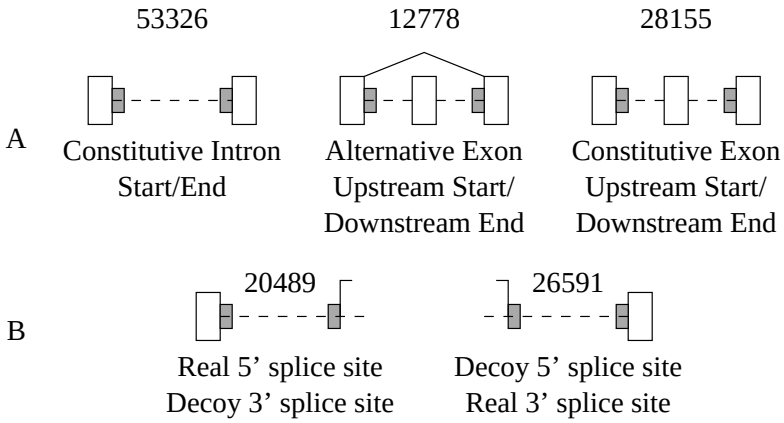


Figure 3-9: 5′ splice site/3′ splice site control sets. White boxes, exons. Gray boxes, locations of motifs. Dashed lines, introns. Angled line, alternative splice. Periscopes, decoy splice sites. Numbers are of sequence pairs in each set. (A) Sequence sets in which the GC/AU motif was found to significantly collocate. (B) Sequence sets in which the GC/AU motif was not found to significantly collocate.

Still, the most interesting evidence came from our control sets. We constructed these sets of pairs of related 5′ splice sites upstream of 3′ splice sites according to different criteria, show in Figure 3-9.

Figure 3-10 shows the significant collocations in four control sets. Starting with the decoy sets, which were constructed by finding high-scoring splice sites within constitutive introns that, according to EST evidence, are never used, we see that there are about as many significant collocations as one might expect by chance given our significance cutoff of one expected false positive per set.

More interestingly, we find the GC/AU motif pair in distal splice sites of introns flanking both alternative and constitutive exons. The lack of any significant collocations at $k = 6$ probably reflects a decrease in statistical power for longer k-mers—the average counts per k-mer pair goes down by a factor of 16 for each increment of k.

The finding of the GC/AU motif pair in distal splice sites of the alternative and constitutive exon sets suggests that it regulates the inclusion of intervening exons. However, the effect of this regulation is at first glance unclear. The pair's collocation distal to constitutive exons suggests it functions to enhance inclusion of the intervening exon, yet its collocation distal to alternative exons suggests suggests that it functions to suppress intervening exons. Interestingly, no more significant GC/AU collocations were detected in the constitutive exon set than in the alternative exon

Figure 3-10: Significant collocations in 5′ splice site/3′ splice site control sets. Collocations for $k = 4, 5$ and 6 are shown in the left, middle and right columns. Within each cell the k-mers on the left occur in the first 80 nucleotides after 5′ splice site and the k-mer on the right occur in the last 80 nucleotides before the 3′ splice site of the corresponding control set (shown at far right). The number of sequence pairs in each set is shown. Set diagrams as in Figure 3-9.

set for both $k = 4$ (31 for constitutive and 7 for alternative) and $k = 5$ (7 for both sets), and yet there were over twice as many sequence pairs in the constitutive set ($N = 28155$) than the alternative set ($N = 12778$). Therefore, the signal of this collocation in the alternative set is much stronger than in the constitutive set. These data suggest that the GC/AU motif pair primarily functions to suppress intervening exons. This interpretation is also supported by the collocation in the constitutive intron set in which it was first discovered. Its presence in the constitutive exon set may reflect contamination of alternative exons whose skipping isoforms are not represented in the EST databases.

3.2.6 GC/AU Motif Pair Enriched in Long Introns

We next set out to analyze the length distribution of constitutive introns containing the GC/AU motif pair. To simplify the analysis we defined the GC motif as a stretch of seven nucleotides containing at most one A or T. Likewise we defined the AT motif as a stretch of seven nucleotides containing at most one G or C. Then, for each of the 53326 constitutive introns in our starting set we counted the number of non-overlapping copies of each motif it contained, and plotted empirical cumulative distribution functions (ECDFs) of the length distributions conditional on different motif copy numbers.

Figure 3-11 shows these plots. Looking at the plots from top to bottom, we first see a slight shift toward shorter introns for multiple copies of the GC motif near the $5'$ splice site. This is not a surprise as short intron length is known to be associated with G+C isochores, for reasons that are unclear. Similarly, we see a shift towards slightly longer introns for multiple copies of the AU motif near the $3'$ splice site. Again, long introns are associated with more A+U-rich regions of the genome. However, when we look at the set of introns that have many copies of *both* motifs at the corresponding splice sites, the shift towards long introns is impressive. Beginning with four copies the shift is clear, but with at least six copies the shift is even stronger. Note the logarithmic scale for the x-axis. This shift represents a roughly 10-fold increase in intron length.

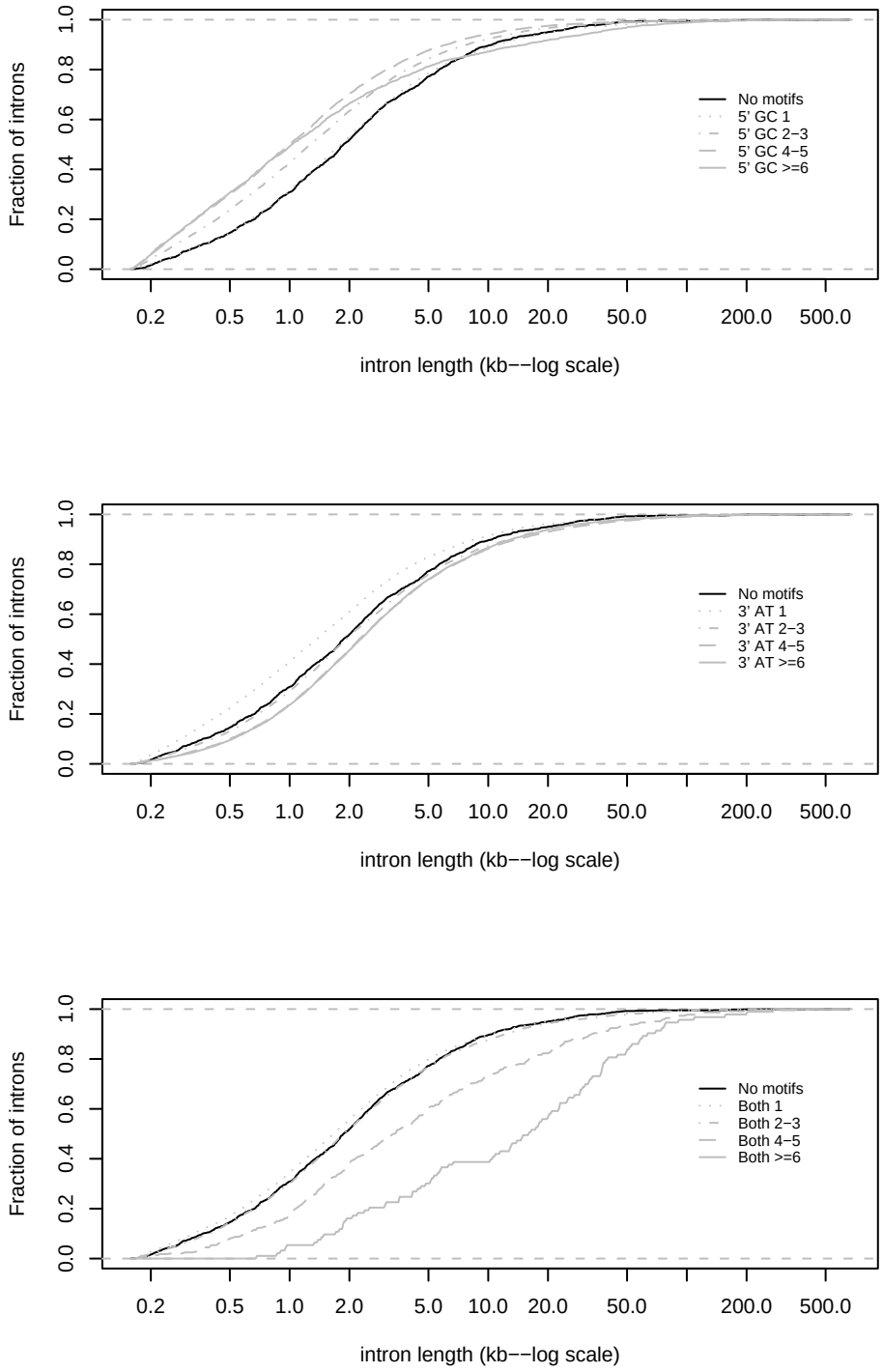


Figure 3-11: Introns with many GC/AU motif pairs tend to be very long. ECDF of lengths of constitutive intron with increasing numbers of GC motif near the 5′ splice site (top), AT motifs near the 3′ splice site (middle), or both motifs at their corresponding splice sites. Note the supershift in intron length for at least 6 copies of both motifs.

As a control, we made the same plots but with the GC motif at the 3′ splice site and the AU motif at the 5′ splice site. The result is shown in Figure A-1 of Appendix A. The GC and AU plots look similar, with a high copy number-dependent shift to short and long introns respectively. As for introns with many copies of both motifs, there are generally fewer when the motifs are flipped, and the shift to long introns is either very weak or absent. This control is important because it rules out the the explanation that the length shift is due to isochore boundaries within introns.

This result argues for a function of the motif pair either in promotion of long splices or, since the number of decoy splice sites in an intron is generally proportional to its length, suppression of intervening exons.

3.2.7 The GC/AU Motif Pair Suppresses the Middle Exon of a Three Exon Mini-Gene*

In order to probe the effect of the GC/AU motif on an intervening exon we used a three exon mini-gene construct originally created by Dr. Zefeng Wang (Figure 3-12). When the middle exon is skipped exons 1 and 3 encode a functional Green Fluorescent Protein, whereas the inclusion of the middle exon, which is from the IMP1 gene [29] yields a non-functional protein. We cloned the first motif after the HindIII site 16 nucleotides from the 5′ splice site to clone the first motif, so that it begins 22 bases after the 5′ splice site, and the second motif 17 bases upstream of the SacII site which itself is 14 bases upstream of the 3′ splice site, so that the second motif ends 37 bases upstream of the 3′ splice site. The reason for the extra spacer was to avoid making our insertion between the branch point and polypyrimidine tract.

Three representative daydreamers were inserted at the two insertion points. The representative for the GC motif was GGGCGCGGGCGC, and for the AU motif the representative was TTAAAATTAAAA. These are duplicates of the most significant hexamer collocation for the constitutive intron start/end set. The third motif was CGGTTACGAGTA, a neutral motif chosen to avoid known splicing regulatory elements and composed

*Mr. Daniel Ding and I performed these experiments with the help of Drs. Noam Shomron and Zefeng Wang.

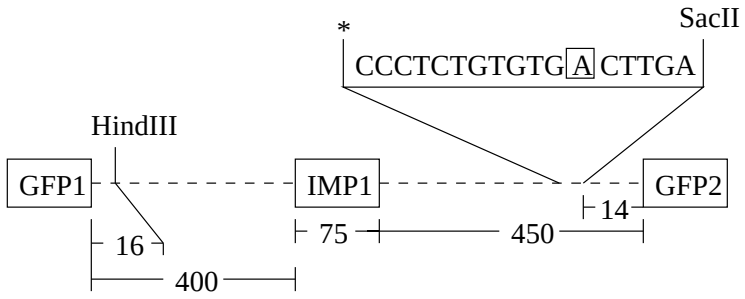


Figure 3-12: Mini-gene construct for interrogating GC/AU motif pair. A , putative branch site. *, second insertion site (first insertion site immediately follows HindIII site). Both restriction sites are six nucleotides long (not shown). Nucleotide lengths indicated.

exactly 50% of G+C and 50% of A+T (see Section 3.4.4).

Five constructs were made as follows:

Construct	1	2	3	4	5
5′ splice site motif	Neutral	GC	Neutral	GC	AU
3′ splice site motif	Neutral	Neutral	AU	AU	GC

These five constructs were transiently transfected into HeLa cells and their splicing assayed by quantitative RT-PCR (Figure 3-13). Construct 4 (GC motif at the 5′ splice site and AU motif at the 3′ splice site) showed a roughly 2 fold increase in skipping frequency relative to the neutral construct, significant at the $P \leq 5^{-4}$ level by a one-sided t-test. None of the other constructs exhibited differences in skipping levels, confirming our prediction that these motifs function only as a pair to suppress the inclusion of intervening exons.

3.3 Discussion

We presented a statistical method for the detection of functional motifs in paired sequence sets. The method successfully controls for the known G+C-heterogeneity of the human genome and, in the case of intron 5′/3′ ends rediscovers the U12 splice sites which comprise less than 0.2% of all introns [5]. This method also revealed a novel GC/AU motif pair which functions to suppress the inclusion of an intervening exon in a mini-gene construct.

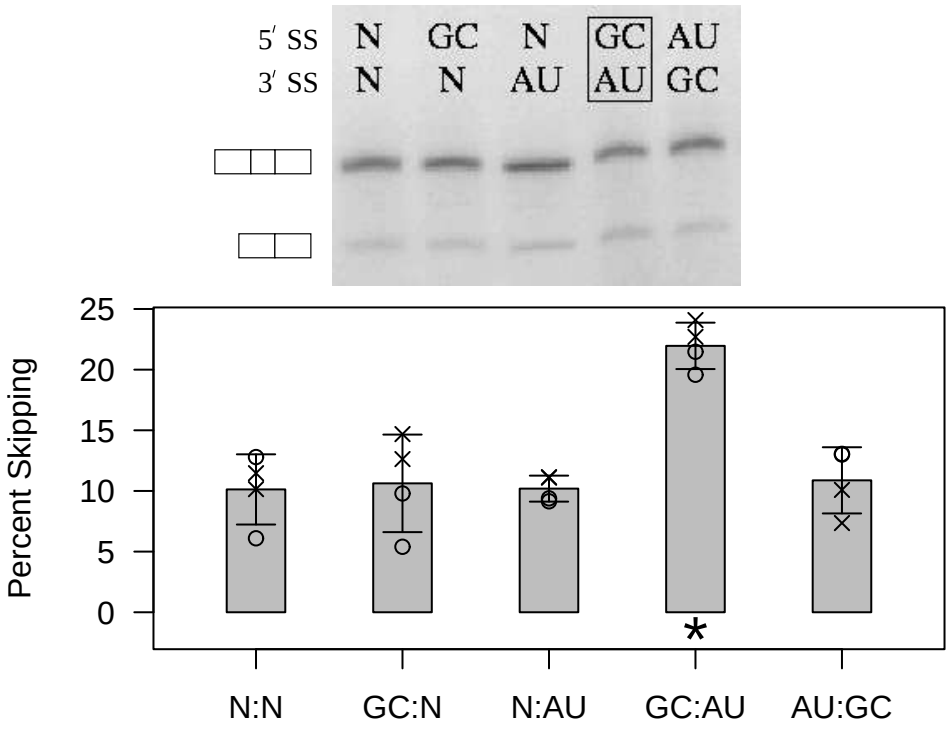


Figure 3-13: `GC/AU` motif pair promotes exon skipping in HeLa cells. Five constructs (see text) were transfected into HeLa cells. Twenty-four hours later RNA was extracted and quantiative RT-PCR was performed (20 cycles of PCR) to assay for relative isoform levels. Top, representative gel showing levels of inclusion isoform (upper band) and skipping isoform (lower band). Motifs inserted at $5'$ and $3'$ splice sites are shown above each lane. Box indicates `GC/AU` motif pair. Bottom, quantitation of skipping levels. Five different constructs are shown on the x-axis with $5'$ splice site and $3'$ splice site motif inserts separated by colons. `X`'s and `O`'s represent repetitions of quantitative PCR for two independent transfections. Bar heights and errors are means and standard deviations of four data. `*` indicates significance at the $P \leq 5 \times 10^{-4}$ level by a one-sided t-test versus N:N construct. No other constructs are significantly different from N:N even at the 5% level.

3.3.1 Stratification Controls for G+C Heterogeneity

The G+C heterogeneity of the mammalian genome was first observed in the 1970s by separating genomic fragments in a cesium chloride gradient (reviewed in [17]). The genome is comprised of long regions within which the G+C content is homogeneous, but between which the G+C content varies extensively. The homogeneity is so strong that G+C content of splice sites, introns, and third-codon positions all correlate with the G+C content of the region in which they lie. Several theories have been put forth to explain this heterogeneity but no clear consensus has yet emerged. Whatever the reason for G+C heterogeneity, it will certainly cause k-mers of similar G+C content to apparently collocate, and could potentially mislead an investigator (for example Figure 3-2).

Our stratification method successfully controls for this effect, as no significant collocations were detected in co-GC shuffled intron ends (Figure 3-7). The method works because the null hypothesis takes into account the expected number of collocations in each intron based on the G+C content of its ends. A simplification of this method which we have not tried is to stratify only on the G+C content of the entire intron rather than in both dimensions on the co-G+C content. We expect that this variation will give similar results.

3.3.2 Humans Likely Have No More Than Two Spliceosomes

Given the ability of our method to detect the U12 splice sites, even though they account for only 0.2% of human introns, our results indicate that a third spliceosome does not exist. There are, however, a number of scenarios in which a third spliceosome could escape notice by our methods. First, this third type of intron might be less abundant than the U12 introns. This would mean that there could be at most about 100 such introns in all the human genome.

A second scenario is that the splice sites would be more degenerate than the U12 splice sites. Then no collocations would be detected because the counts of any particular hexamer would be depleted by the degenerate positions. We consider this

possibility unlikely because of the general observation that degeneracy of splice sites is negatively correlated with the number of introns spliced by a spliceosome (consider yeast and U12-type splice sites as non-degenerate and intron-poor and mammalian U2 splice sites as degenerate but intron-rich). Still, it is a possibility we cannot exclude.

Finally, if the splice sites were very long (longer than 6) we might not detect them if the hexamers making up these longer splice sites were relatively abundant. In that case, the noise of the hexamers appearing outside of these hypothetical splice sites would mask out the signal of the hexamers collocating within the splice sites.

Thus our data strongly suggest but do not prove that humans lack a third spliceosome.

3.3.3 Suppression of Pseudoexons

With human introns on average about 10 times longer than human exons, how is it that pairs of decoy splice sites appearing within introns ("pseudoexons") don't tend to confuse the spliceosome? We have identified one possible mechanism by which cells can avoid this error. The GC/AU motif pair, with a G+C-rich motif at the beginning of an intron and an A+U-rich motif at the end, can function to inhibit the inclusion of intervening exons. These motifs collocate across both constitutive introns and distal splice sites of introns flanking alternatively spliced exons (Figures 3-8 and 3-9), they are found in high copy number at the ends of very long introns (Figure 3-11), and they function in a mini-gene construct to suppress the inclusion of an intervening exon (Figure 3-13). Thus the GC/AU motif can potentially be a mechanism for cells to suppress pseudoexons.

In order to estimate the importance of this motif pair, we counted the total number of collocations in the intron $5'/3'$ end set of any of the G+C-rich pentamers in the GC/AU cluster at the $5'$ end and any of the A+T-rich pentamers in the GC/AU cluster at the $3'$ end. We found 3365 such collocations, suggesting that the motif pair might be involved in the splicing of as many as 6% of the 53,326 constitutive introns in our set. In the case of distal splice sites of introns flanking skipped exons, we find such 1126 collocations, suggesting that the motif pair might be involved in the splicing

of as many as 9% of all 12,778 skipped exons in our set. The GC/AU motif pairs are widespread in both sets, and if functional in these contexts are therefore a key element in splicing specificity.

3.3.4 Motifs from Collocating k-mers

One problem which did not arise in these sequence sets, but is in general an important issue, is the problem of grouping together k-mers into functionally relevant clusters. In my opinion the best method for doing this would ignore the actual sequence similarities of the k-mers. This would come into play as a control after the analysis was finished, or could be added independently.

Given a set of k-mers X appearing in the first sequences and another set Y appearing in the second sequences, how can we ask if the sets as wholes collocate? We can follow the analysis of Section 3.2.3 exactly, replacing "contains hexamer x" with "contains any hexamer $x \in X$" and "contains hexamer y" with "contains any hexamer $y \in Y$". Thus we can calculate a P value for setwise collocations, measuring all the pairwise interactions between the sets simultaneously.

There are two central problems to be solved. The first is determining which sets X and Y to look at (there are $(2^{4^6} - 1)^2 \approx 10^{2466}$ pairs of such sets of hexamers!), and the second is the calculation of statistical significance, given the enormous space we are searching.

A possible solution to the first problem is to cluster the *collocations*, that is pairs of hexamers (x, y). Then a set of collocations $\{(x_i, y_i)\}_i$ induces a pair of k-mer sets $\{x_i\}_i$ and $\{y_i\}_i$. Thus the setwise collocation P value could constitute a distance measure to be used for a hierarchical clustering. Thus a tree could be formed with all the collocations at its leaves where the height of an internal node is the P value for setwise collocation of the sets induced by all its leaves. This method also leaves open the possibility of a particular hexamer functioning as part of two different complexes depending on its collocating partner. As described here, it has the problem that the height of a parent might be less than the height of a child. It is a future research direction to sort that out.

3.3.5 A New Type of Motif Finder

Traditional motif finders such as Gibbs Sampler [31] or MEME [1] search for statistically enriched motifs in sequence sets, assuming that these motifs should exhibit biological function. The basis of this assumption is that neutrally evolving sequences should have a steady-state k-mer distribution depending only on the parameters of evolution of the organism's DNA [25], which are controlled for either by appropriate null-hypotheses, sequence shuffling, or with a negative control sequence set (e.g. see Section 4.2.4).

Here we solve the problem of motif pair finding by looking for pairs of motifs that are co-enriched. Our null-hypothesis now controls for all k-mer frequencies in each set, regardless of whether those frequencies are due to neutral or purifying selection. When two motifs function in concert, they should exhibit a different selective pressure together than separately, and hence their occurrences should not be independent, as we would expect from non-interacting motifs. Thus the basic principle of traditional motif finding, "enrichment implies function", is also the basic principle of our paired motif finder.

Furthermore, just as traditional motif finders come in different flavors that can solve variants of the motif finding problem, so could many flavors of paired motif finders be created. One variety that would be particularly useful would identify functional motif pairs in unpaired sequences. Such a program could, for example, detect pairs of functionally interacting miRNAs by analyzing the co-distributions of their seed matches in $3'$ UTRs. Another useful flavor would allow for detection of degenerate motifs (see Section 3.3.4). We expect that the basic ideas laid out in this manuscript could be applied to these and other similar problems with only a few careful changes.

3.4 Materials and Methods

3.4.1 Sequence Sets

The sequence sets used in this analysis all came from the program SpliceGraph [57] developed by Drs. Rickard Sandberg and Michael Stadler. SpliceGraph downloads genomic alignments of ESTs from the UCSC website `http://genome.ucsc.edu` [28] and calculates the coordinates of all the splice sites necessary to generate each set. It next clusters ESTs that share splice sites into gene models. Finally, it parses the gene models to define sets of splicing variation patterns such as constitutive exon or alternative exon.

The sequence sets were all prepared by Dr. Stadler. In all cases he used the beginning and ending 80 nucleotides of the region.

The *Constitutive Intron Start/End* set was made from constitutive introns in SpliceGraph that were at least 160 nucleotides long (so that the two regions did not overlap). These were defined as pairs of splice sites that were supported by EST evidence and for which, furthermore, there was no evidence of any alternative splicing in that region.

The *Alternative Exon Upstream Start/Downstream End* set (also referred to in the text as "distal splice sites of alternative exons") was made from SpliceGraph Exons and flanking introns for which there was evidence of both inclusion and exclusion. Furthermore, the introns were both required to be at least 160 nucleotides long, so that the set could be used without resulting in overlapping 80-mers for analyses of the proximal splice sites, which are not discussed here. The exons were also required to be at least 80 nucleotides so that 40 nucleotides of non-overlapping exonic sequence could be extracted near each splice site for another analysis that was not discussed here.

The *Constitutive Exon Upstream Start/Downstream End* set (also referred to as "distal splice sites of constitutive exons") was made from constitutive exons at least 80 nucleotides long (again due to the constraints of another analysis not discussed here) with flanking constitutive introns that were both at least 160 nucleotides long.

The *Real* 5′ *Splice Site/Decoy* 3′ *Splice Site* set was made by taking the Constitutive Intron set and searching for decoy 3′ splice sites at least 160 nucleotides from the 5′ splice site. A decoy 3′ splice site was defined as a stretch of 23 nucleotides which scored at least as high as the 20 intronic and 3 exonic nucleotides at the true 3′ splice site in the Maximum Entropy model [64]. The *Decoy* 5′ *Splice Site/Real* 3′ *Splice Site* set was formed analogously.

3.4.2 Sequence Set Filtering

Two steps were performed on each sequence set. First, Dr. Stadler removed any sequences that overlapped with annotated repeats (also from the UCSC website [28]). These included both repetitive elements such as Alus and LINEs and shorter simple sequence repeats.

In the second step we removed paralogous sequence pairs. To do this, three similarity graphs were made for each set. These graphs had the same node set, one node for each sequence pair in the set. The first graph had an edge for each significant BLASTN hit of nucleotides +10 to +85 from the 5′ splice site, and the second graph the same for nucleotides -100 to -25 from the 3′ splice site. These regions were chosen to avoid the most information-rich parts of the splice site. The third graph had as its edge set the intersection of the first two graphs, so that two sequence pairs were connected if and only if *both* regions showed sequence homology.

From this intersection graph we performed greedy node removal, iteratively removing the node of highest degree (and any attached edges) until no edges remained, so that there were no two sequence pairs with both ends homologous. The sequence pairs corresponding to the remaining nodes formed the input to the collocation algorithms.

3.4.3 Co-GC Shuffling

Our implementation of co-GC shuffling differed from the simplified description in the caption of Figure 3-4. In fact we generated the co-GC shuffled sets in "one fell swoop"

without any swaps. As we walked through the real sequence pair sets, whenever we saw a sequence pair of co-GC content (s_1, s_2) we chose at random, and without replacement, a first sequence (i.e. first in its pair) of co-GC content s_1 and, independently, a second sequence of co-GC content s_2. These together formed one co-GC shuffled sequence pair. Thus the final set had the same overall co-GC content but very few, if any, of the original pairs.

3.4.4 Selection of the Neutral Motif

The neutral motif used in the experiments was chosen so that, when inserted into either cloning site, it would not create any known splicing regulatory elements. Our set of "known" splicing regulatory elements was the union of the RESCUE-ESEs [15], the cut2 FAS-ESS hexamers [61] and a set of putative ISE and ISS hexamers provided by Dr. Grace Xiao determined from her computational screen. The motif does not contain any of the hexamers, nor does it overlap any at either of the two contexts. Furthermore, it has 50% G+C content.

3.5 Acknowledgments

Dr. Michael Stadler collaborated with me for much of the computational work in this study. He created the sets of sequences based on SpliceGraph, filtered out annotated repeats, helped with an implementation of the algorithm and some data visualization, and is generally an invaluable colleague and friend.

Dr. Zefeng Wang provided the initial mini-gene construct and helped with some of the experiments, and Dr. Noam Shomron oversaw the rest of the experiments, even performing some himself. Mr. Daniel Ding spent many hours in the laboratory creating the five different constructs and working out the quantitative PCR protocol.

Chapter 4

A Sequence Model for the Branch Site

Abstract

The first catalytic step of splicing is the formation of the branch point, a special adenosine about 25 nucleotides from the end of the intron. This adenosine and its flanking recognition sequence are necessary for the splicing reaction to proceed, and mutations in these positions are associated with a number of different human diseases. Nevertheless, the branch point has been poorly characterized and in fact only a handful of natural branch points have been accurately mapped. We built a sequence model of intron 3′ ends that is capable of both scoring potential 3′ splice sites and predicting branch points. The model was trained using an EM algorithm with initial guesses based on introns containing a single adenosine within a restricted region. It was validated on a set of 24 branch points reported in the literature. We used the model to identify a set of nearly 40,000 high-confidence human branch points, which we estimate is 90% accurate. We examined the conservation of branch point position and found that it is highly conserved between human and mouse. Finally, we applied the model to study the role of the branch point in alternative splicing, and found that skipped exons show an enrichment for misordered 3′ end elements relative to constitutive exons, and that although branch point sequences of skipped exons are

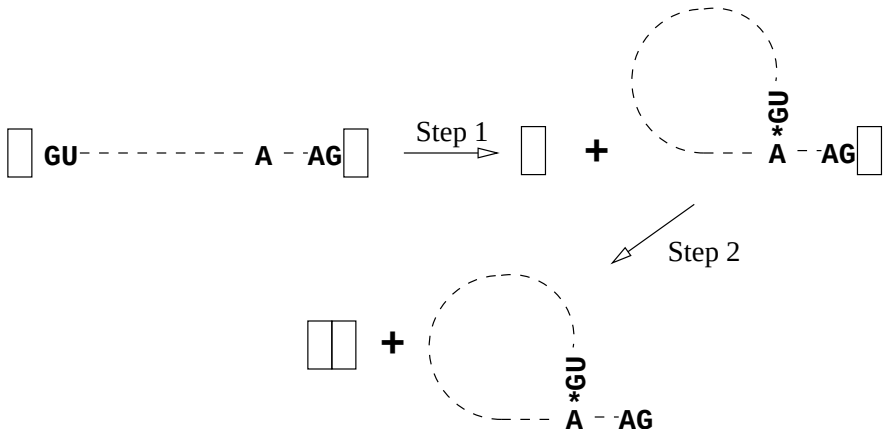


Figure 4-1: The two catalytic steps of splicing. Step 1, 2' OH from branch point adenosine attacks 5′ phosphate of guanosine at beginning of intron (*), leaving upstream exon with 3′ OH and covalently linked lariat and downstream exon. Step 2, free 3′ OH attacks 3′ phosphate of guanosine at end of intron, liberating lariat and ligated exons.

more highly conserved than those of constitutive exons, their positions are conserved at the same rate.

4.1 Introduction

The RNA splicing reaction occurs in two catalytic steps (Figure 4-1). In the first step a specific adenosine is recognized, usually near the -25 position of the intron (i.e. 25 bases 5′ of the first base of the downstream exon). This base is thought to bulge out of an RNA helix formed between the intron and the U2 snRNA, although it is initially recognized as part of a complex involving the protein SF1 (also known as branch point binding protein, BBP), and stabilized by a nearby polypyrimidine tract and the protein U2AF65. The 2' OH group of the sugar attacks the 5′ phosphate group of the +1 guanosine. The final products, the ligated exons and an excised lariat, result from the second step of splicing [6].

The branch point is essential for the reaction, which often will not go forward if it is mutated. Indeed, the first step of splicing can occur in the absence of a 3′ splice site `AG` (acceptor site), but not in the absence of a branch point [44]. Furthermore, the choice of 3′ splice site is thought to be determined by scanning downstream for the first `AG` beginning about 10 bases after the branch point [52], [54].

The branch point is also known to be important for alternative splicing of the α-tropomyosin and α-actinin genes ([53], [55] and see Figure 1-9). These genes both contain pairs of exons whose mutual exclusion depends on the position of the branch point upstream of the second exon of the pair. This branch point is located so close to the first exon of the pair that the formation of the lariat is sterically hindered. This was proved by insertion of a spacer between the first exon's $5'$ splice site and the branch point, which allowed inclusion of both exons.

Despite their importance, human branch points have not been extensively studied. A recent literature survey found only 19 mapped mammalian branch points [29]. One reason for the paucity of known branch points is the difficulty in their experimental determination. Branch points can be mapped either by genetic or biochemical methods. For the genetic method one needs to have an idea of which base is the branched adenosine, then mutate it and show that splicing is abolished. This method may not always work because some introns can splice through more than one branch point. The result can also be questioned, as mutation of an intronic splicing enhancer could yield the same phenotype.

There are several biochemical methods to map branch points, but most require large amounts of lariat RNA, which is not abundant in mammalian cells. Therefore, it is necessary to create a mini-gene and to splice *in vitro*. This has the problem of physiological relevance. Once lariat RNA has been isolated, it can be debranched into linear intron using the enzyme debranchase. The branch point can be mapped by looking at differential mobility of an RNAse fragment in the two samples, by looking for a lariat-specific reverse transcriptase (RT) stop, or by RT-PCR with primers across the branch point. The latter two methods depend on the pausing of certain reverse transcriptases at the branched adenosine. This yields a stop in the primer extension assay, but also allows for cross-lariat amplification in the RT-PCR assay.

Previously, Kol et al computationally modeled branch sites and studied the difference between alternatively spliced and constitutively included exons in human and mouse [29]. They built a position specific scoring matrix for the branch site based on 19 experimentally proven branch points, and used it to predict branch sites in authen-

tic introns by searching for high-scoring branch site sequences followed by a tract of at least 14 nucleotides of at least 50% pyrimidine. Using this model they found that the branch site sequences of alternatively spliced exons are more conserved between human and mouse than those of constitutively spliced exons. They also mutated their predicted branch sites using RT-PCR of mini-gene constructs and determined that mutating the branch point adenosine or position -2 to guanosine increased exon skipping.

Knowing the difficulties of experimental mapping, we decided to create a computational model of intron 3′ ends, including the branch point, the polypyrimidine tract, the `AG`, and the first three bases of the exon. Our model differs from the Kol model in that it uses a larger training set, an EM algorithm to improve the initial training guesses, and gives a posterior probability indicating the confidence of any given branch site prediction. We applied our model to study differences between alternatively and constitutively spliced exons.

4.2 Results

4.2.1 A Linear Sequence Model for Intron 3′ Ends

We developed a Linear Sequence Model (see Section 4.4.1) for the 3′ end of introns. We used seven ordered states (Figure 4-2, BPA): intronic, branch point site (BPS), first gap, polypyrimidine tract (PPT), second gap, acceptor site (the last three positions of the intron, consensus `YAG` [`Y` = `C` or `U`]), exonic. We call this the BPA model because of the three conserved elements: BPS, PPT and acceptor site.

The BPA model can make two types of predictions. The first type of prediction is that, given a sequence extending three bases into the beginning of an exon, it can generate a score for the whole region. A higher score indicates that the region contains all the elements necessary for recognition as a 3′ splice site by the spliceosome, and a lower score indicates that the region looks more like internal intron sequence. Thus the model can be used to help discriminate decoy and real 3′ splice sites. This is the

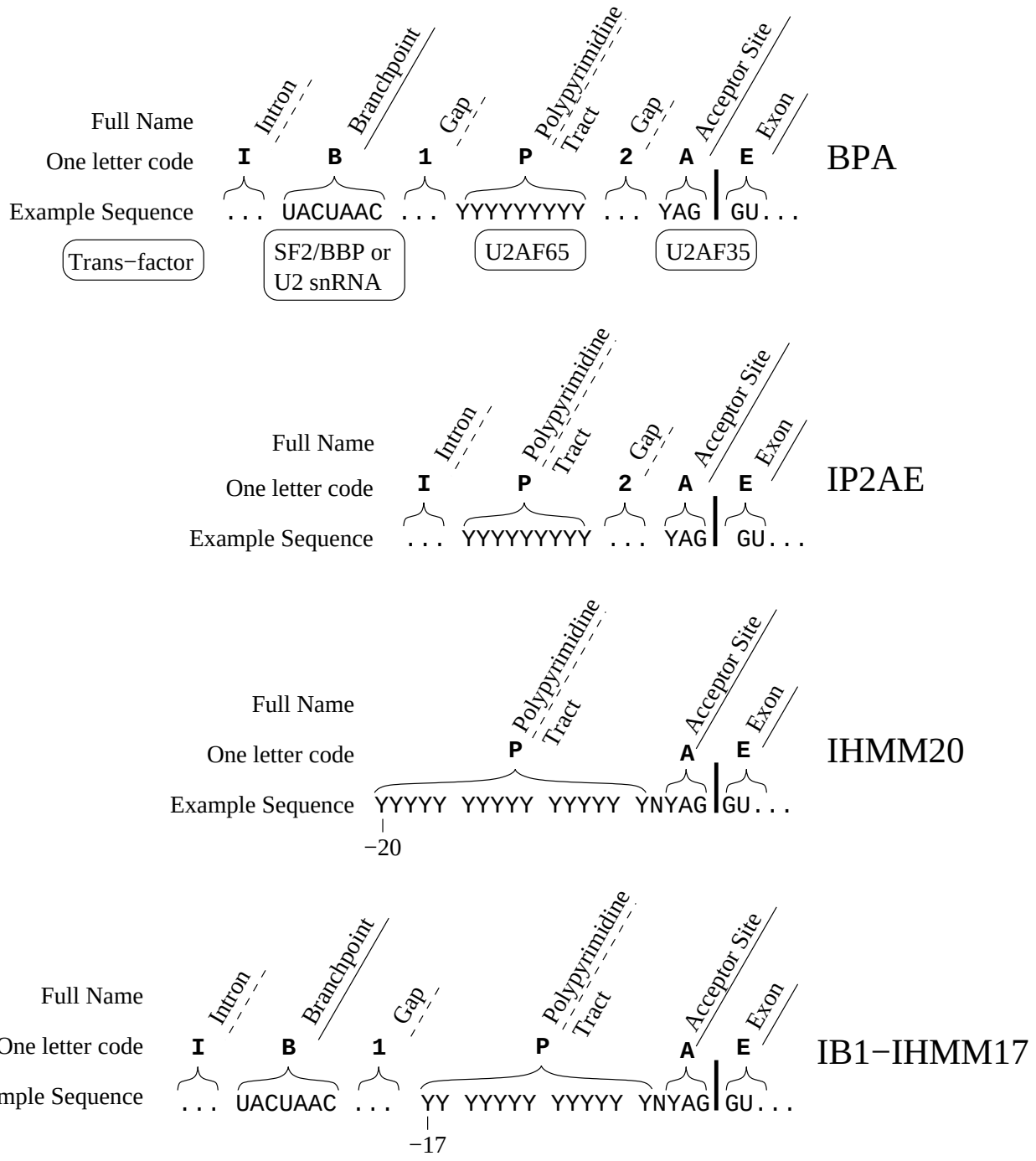


Figure 4-2: The structure of the BPA linear sequence model and related models. States of fixed length are shown with solid underline and of variable length with dashed underline. The boundary between the intron and the downstream exon is shown as a thick vertical line. Known binding trans-factors (for BPA only) and typical sequences (Y = C or U) are shown for the conserved elements.

same problem that is addressed by the Maximum Entropy model for 3′ splice sites [64]. As an example, Figure 4-3 shows the model score for every AG in the full length transcript of SirT1.

The second type of prediction is that, given an authentic 3′ splice site, it can generate posterior probabilities for the BPS. This means assigning one probability p_i for each position i in the sequence with $\sum_i p_i = 1$. This probability can be interpreted as the likelihood that that position is the true branch point. Figure 4-4 shows these distributions for the adenovirus major late transcript (AdML), where the branch point has been experimentally mapped and correctly predicted, and the 8 real 3′ splice sites in SirT1, whose true branch points have not been mapped.

4.2.2 Creating the BPA Model

Initial Guesses for Emission and Length Parameters

The examples of the last section were created with our fully-trained BPA model. To create this model, we began with an "educated guess" for all parameter values except for the branch point state, which was trained on a set of 80 introns we generated that contained only a single adenosine in the region beginning 40 nucleotides downstream of the 5′ splice site and ending 10 nucleotides upstream. The reason for excluding the beginning and end of the intron is that branch sites too close to the splice sites are thought to be prevented by steric hindrance. Evidence for this is found in the mechanism enforcing the mutual exclusivity of α-tropomyosin exons (Figure 1-9). Experiments involving the insertion of spacer elements between the 5′ splice site of the first mutually exclusive exon and the branch site of the second show that branch sites cannot form closer than about 50 nucleotides from the 5′ splice site [53]. We chose a conservative cutoff of 40 nucleotides. The evidence for excluding the last 10 nucleotides of the intron comes from experiments introducing AG dinucleotides by mutation at varying distances from the branch site. In general the first AG at least 12 nucleotides downstream of the branch site become the 3′ splice site [52], and AG dinucleotides closer than that are not competitive as 3′ splice sites. We chose the

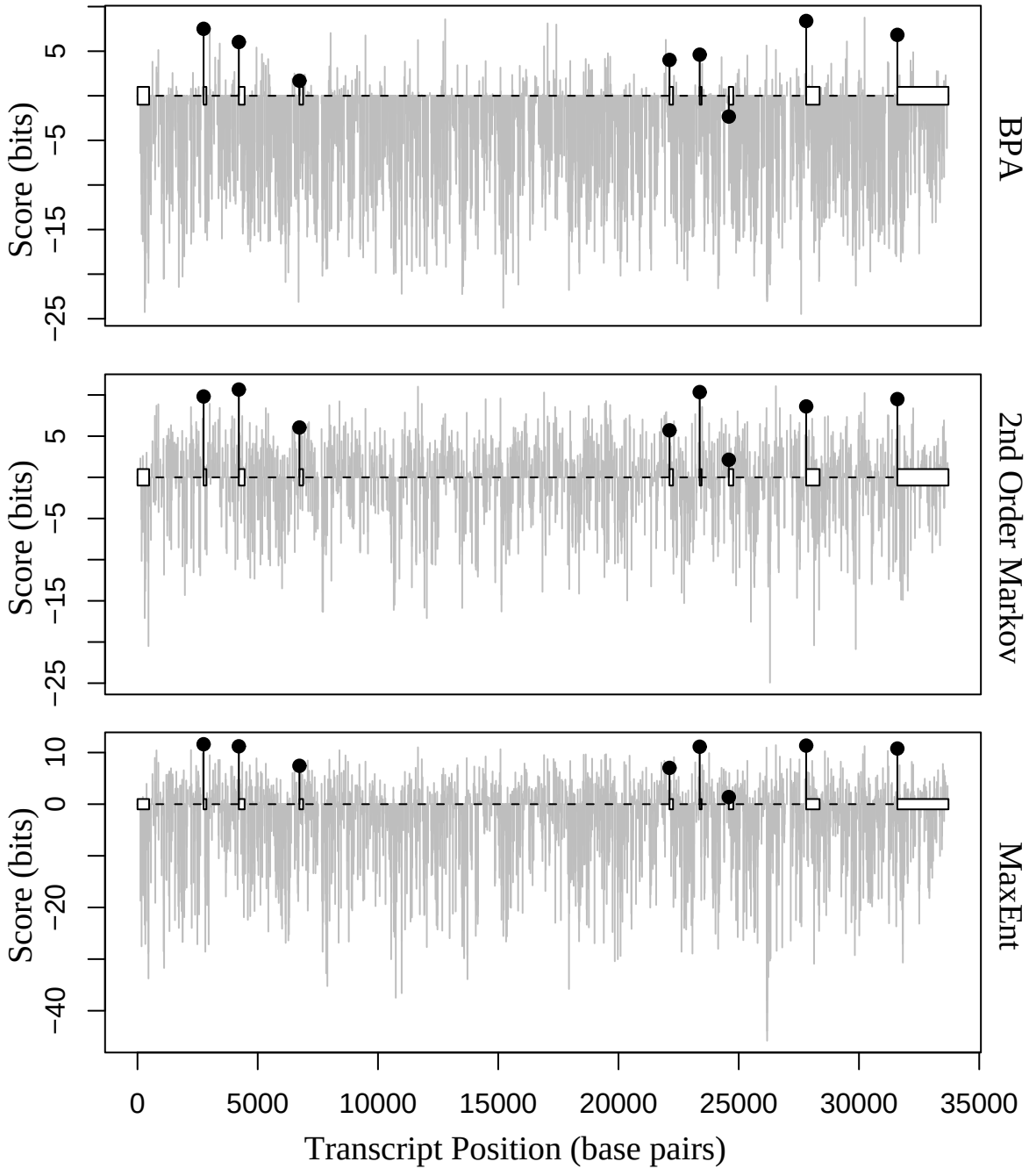


Figure 4-3: BPA scores putative 3′ splice sites. Each bar indicates the score of a potential splice site (`AG`) at that position in the transcript. Exons are shown with white boxes, real 3′ splice sites with black lollipops. Scores are from the BPA model (top), a second order Markov model of the -20..+3 region (middle), or the MaxEnt model [64] (bottom).

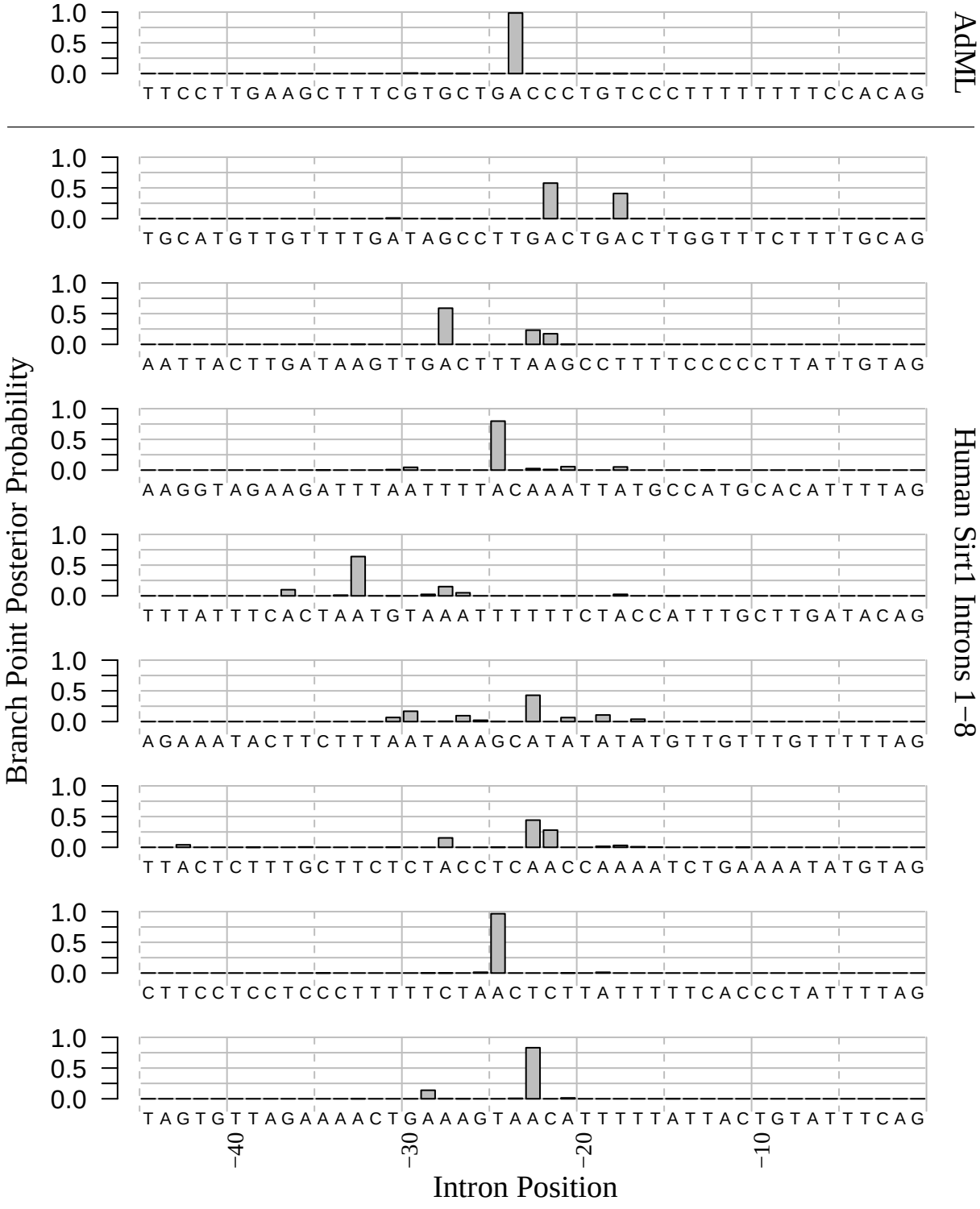


Figure 4-4: BPA calculates posterior probabilities for branch point position

conservative cutoff of 10 nucleotides. Finally, the rule for a single adenosine in the allowed region is based on the fact that nearly all branch sites are adenosines. In fact there are examples where other nucleotides can be used as branch points, but only in the absence of a nearby adenosine [21].

The branch sequences of 80 single-adenosine introns we found were then used to build a positional weight matrix which was the initial guess for the emission probabilities of seven branch point states of our LSM. We guessed length distributions and other emission probabilities based on our intuition from looking at many authentic intron sequences in the expectation that any reasonable guess would be properly trained by the EM algorithm.

Training and Validating the BPA Model

The guesses of the previous section were then expanded into higher order emission probabilities to allow for neighboring dependencies. One or more iterations of the Baum-Welch algorithm were run starting with this guess on a set of about 10,000 real 3′ splice sites taken from SpliceGraph [57], a database of known gene isoforms derived from EST/genomic alignment (see section 3.4.1).

We tried using different orders for the Markov models (0 through 4) and also different number of training iterations (0, corresponding to the guess without training, through 5 iterations). We evaluated the quality of the different models by using the model to score 3′ splice sites within the splicing simulator ExonScan [61]. ExonScan normally uses Maximum Entropy models [64] for the 5′ and 3′ splice sites (MaxEnt), as well as adding bonus scores to putative exons that contain exonic splicing enhancers (RESCUE-ESEs [15]) or that are flanked by the intronic splicing enhancer `GGG`, and penalizing those that contain exonic splicing silencers (FAS-HEX2 [61]). ExonScan performance is measured by Exact Accuracy on a set of 1820 constitutively spliced transcripts, defined as the fraction of exons with both splice sites correctly predicted at a score cutoff with equal numbers of false positive and false negatives.

Figure 4-5 shows the performance of our 30 different versions of BPA when replacing the MaxEnt model of the 3′ splice site in ExonScan. The best one, Markov

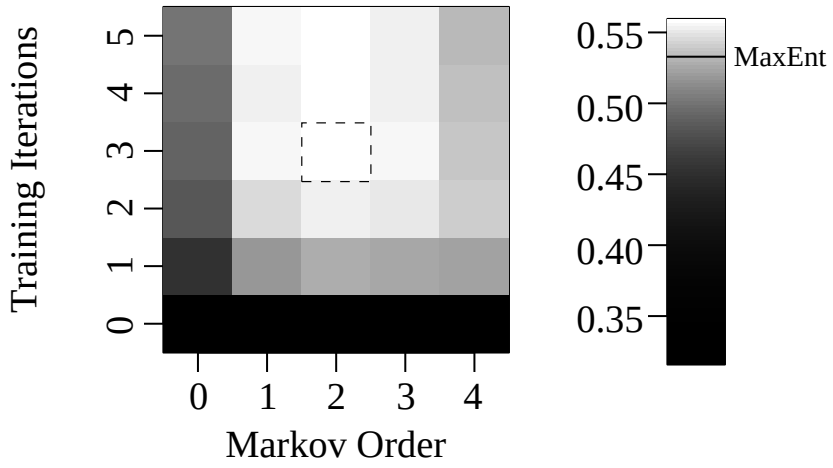


Figure 4-5: Dependence of ExonScan exact accuracy on training iterations and Markov Order when using BPA as the 3′ splice site model. Gray levels correspond to high (white) or low (black) ExonScan exact accuracy, according to legend at right. Dashed box, BPA2.3 model used for subsequent analyses. Black horizontal line on legend, exact accuracy of ExonScan using MaxEnt model as the 3′ splice site model.

order 2 with three rounds of training, achieved an exact accuracy (both splice sites correct) of 55.8% on the test set, compared to 53.3% by MaxEnt. We performed all further analysis with this model, which we call BPA2.3.

Removing the BPS Negatively Impacts BPA2.3′s Accuracy

We next set out to understand the source of BPA2.3′s accuracy in splice site prediction. In particular, we wanted to know if the branch point part of the model was actually useful in predicting 3′ splice sites, or if the PPT/acceptor/exon portion of the model was the reason for its relative success in this task. We formed a shortened model that contained the intronic state followed by PPT, a gap state, acceptor and then exonic states, with parameters taken from BPA2.3 (Figure 4-2, IP2AE). When this model, called IP2AE for the states it contains, was used in ExonScan, it gave an exact accuracy of 46.3%. Three training iterations brought the accuracy up to 49.7%, still short of the full model (55.8%). Therefore, the branch point is important for splice site prediction.

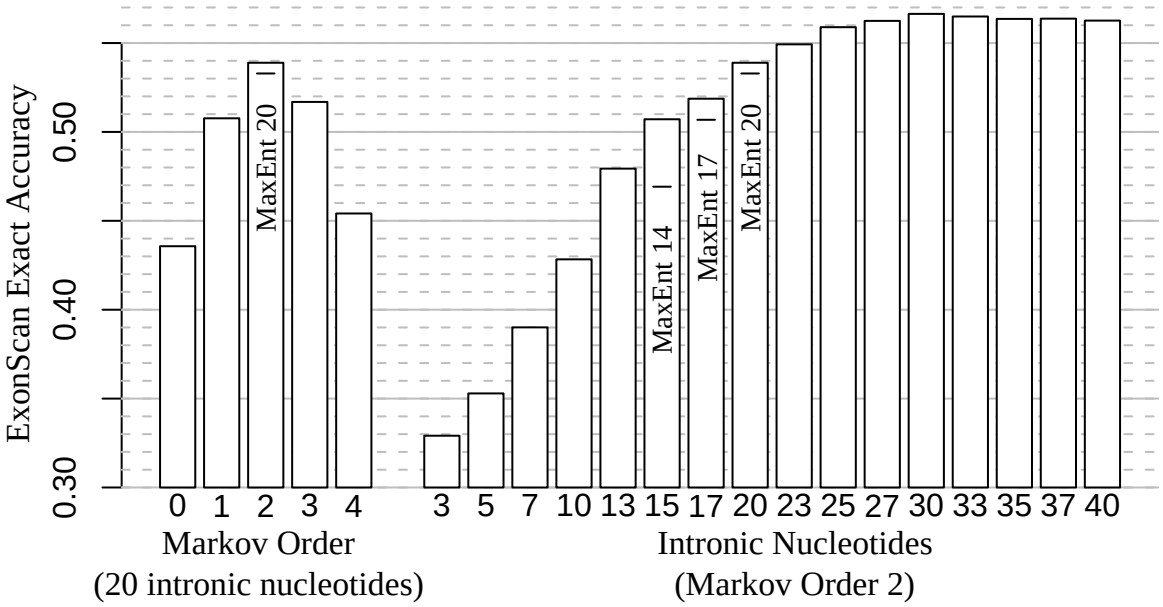


Figure 4-6: Inhomogeneous Markov models (IHMMs) of intron 3′ ends are comparable with MaxEnt in terms of ExonScan Exact Accuracy. Left, dependence of Exact Accuracy on Markov order. Right, dependence on Model length. Horizontal lines indicate accuracy of MaxEnt models of comparable lengths, indicated.

An Inhomogeneous Markov Model Outperforms BPA2.3 and MaxEnt in 3′ Splice Site Prediction

We noticed that the regions modeled by IP2AE and MaxEnt are basically the same, and yet MaxEnt performs better (53.3% versus 49.7%). Therefore, we reasoned that we might be able to improve BPA2.3 in terms of ExonScan exact accuracy if we modeled the polypyrimidine tract similarly to MaxEnt. Since the MaxEnt formulation doesn't fit into our LSM, we used an inhomogeneous Markov model (IHMM) to model the pyrimidine tract. Figure 4-6 shows that the best such models are of second order. Modeling 20 intronic nucleotides (IHMM20, Figure 4-2, IHMM20) achieves 53.9% ExonScan exact accuracy (compare to MaxEnt's 53.3%). The highest accuracy, 56.6% is achieved by IHMM30 (compare to 55.8% for BPA2.3).

Since the IHMM was so successful, we considered replacing the polypyrimidine tract model from our LSM with this simpler model. We formed a model using the Intronic, BPS, and first gap states of BPA2.3 followed by 17 intronic nucleotides of inhomogeneous second order Markov model. The reason for choosing 17 instead of

20 or more was that we wanted to be sure that the inhomogeneous Markov model would not cover branch points. We called this model IB1-IHMM17, again for the states it contains (Figure 4-2, IB1-IHMM17). It achieves an exact accuracy of 56.6% if it is trained for two iterations. This is a slight improvement over IHMM17 alone, but not over IHMM20, which also achieves an exact accuracy of 56.6%. On the other hand, IB1-IHMM20, the model composed of the intronic, BPS and first gap states of BPA2.3 followed by 20 instead of 17 intronic nucleotides of inhomogeneous second order Markov model, only achieves an exact accuracy of 56.2%, which is actually less that IHMM20 alone. Therefore, the addition of the branch point to this model is actually detrimental.

In conclusion, we learned that adding the branch point can have a small effect on accuracy in predicting $3'$ splice sites, depending on the polypyrimidine tract model used. We also discovered that a relatively simple IHMM can outperform MaxEnt in this task.

Prediction of Known Branch Points

In terms of $3'$ splice site prediction, the models IHMM20 and IB1-IHMM17 both outperform BPA2.3, but we decided nevertheless to use BPA2.3 for branch point prediction. Obviously we couldn't use IHMM20 since it does not model the branch point at all. We could have used IB1-IHMM17 but decided that this would be against our original philosophy of modeling the branch point together juxtaposed with the polypyrimidine tract.

Although it is possible for a given intron to have more than one branch point, for simplicity we decided to have BPA2.3 predict at most a single branch site for each intron. That branch site would be the one with highest posterior probability, and the prediction would be made if this probability was sufficiently high.

In order to choose a posterior probability cutoff P we first compiled a set of 24 trusted branch sites from the literature (Table 4.1). All of these were mapped by one the genetic or biochemical methods described above. Most are from human or viral genes, although mTCRB1 is a mouse gene, and AdML and Ftz are artifi-

Gene	Branch Position	Branch Sequence	Reference
HSV-1 LAT	-25	GAGGGAG	[37]
DQB1 intron 3	-21	CACAGAC	[30]
L1CAM	-19	AUCCAAG	[29]
LIPC	-14	CCCCAAU	[29]
FBN2	-15	UUGCAAU	[29]
HEXB	-17	UUGCAAU	[29]
XPC	-24	UACUGAU	[29]
HEMB	-25	CGUUAAU	[29]
TSC2	-18	GCGUGAC	[29]
LCAT	-20	CCCUGAC	[29]
ITGB4	-17	GGCUCAC	[29]
TH	-22	GGCUGAU	[29]
COL5A1	-23	GACUGAU	[29]
mTCRB1	-42	GUCUCAU	[9]
β-globin IVS1	-37	CACUGAC	[29]
γ-globin IVS1	-30	UUCUGAC	[29]
ϵ-globin IVS1	-31	CUCUAAU	[29]
α-globin IVS1	-19	CCCUCAC	[29]
α-globin IVS2	-18	CACUGAC	[29]
CGRP-1 IVS3-4	-36	CACUCAC	[29]
AdML	-24	UGCUGAC	[45]
Ftz	-29	AGCUAAC	[45]
SV40	-18	UUCUAAU	[39]
HGH	-37	ACCCAAG	[29]
LDL-R	-25	CGCUGAU	[63]

Table 4.1: Trusted Branch Sites

cial constructs spliced in HeLa extract. Figure 4-7 shows the posterior probability of these real branch sites and the dependence of the true positive rate (No. correctly predicted BPS/total No. predicted BPS) and coverage rate (total No. of predicted BPS/25) on the cutoff P.

At $P = 0.9$ we predict 5 branch sites of 25 of our trusted set, and all of them are correct, corresponding to a true positive rate of 100% and a coverage rate of 20%. More conservative estimates of these rates come from the smoothed curves (dashed lines in Figure 4-7, right): 96.7% true positive rate and 16.8% coverage rate.

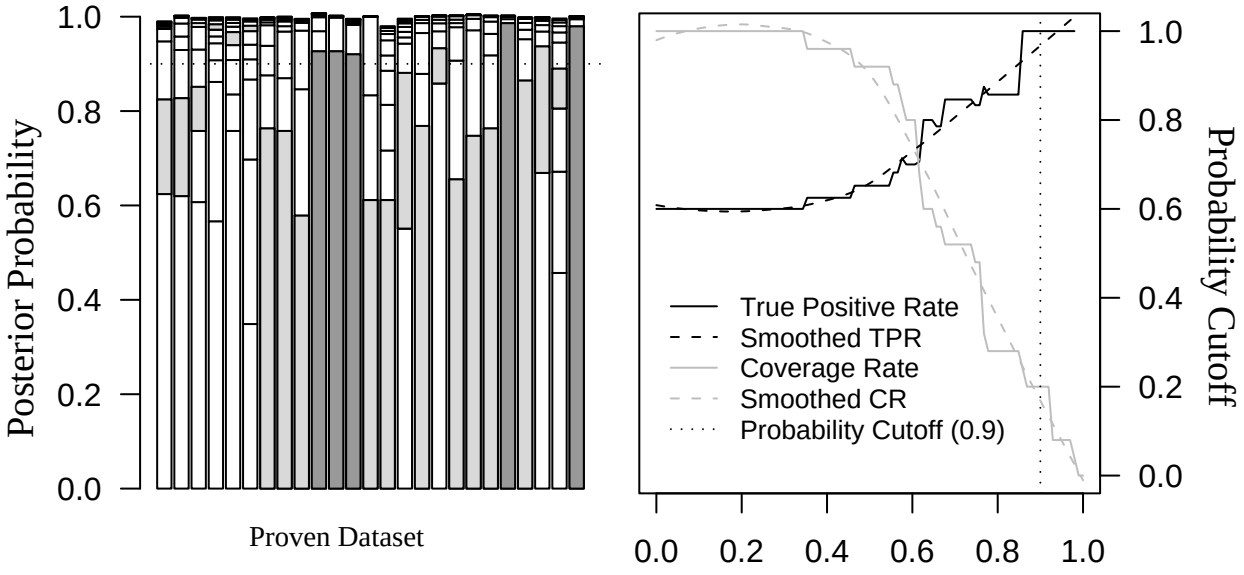


Figure 4-7: Using BPA2.3 posterior probabilities to predict trusted branch sites. Left, posterior probabilities of top 10 positions in each of 25 trusted branch sites. Each column represents one intron, with its top ten posterior probabilities stacked vertically starting with the highest at the bottom. Gray bars, true branch sites. Dotted line, probability cutoff of 0.9. Dark gray bars, predicted branch sites at cutoff 0.9. Right, true positive/coverage curves. Solid lines, real curves. Dashed lines, smoothed. Dotted line, probability cutoff of 0.9.

4.2.3 Prediction of 38,000 Novel Human Branch Points

We next set out to predict as many human branch points as we could. We began with the set of all human 3′ splice sites from SpliceGraph [57] (n = 275,960). Of these 38,017 had a putative branch site with a BPA2.3 posterior probability at least 0.9, giving a coverage rate of 13.7% for this larger dataset. This constitutes our first high-confidence set of branch points, and represents by three orders of magnitude the largest such set ever constructed. Although we expect the set to exhibit some biases (for example low adenosine content) due to its construction, it has roughly the expected sequence composition and distances to 3′ splice site (Figure 4-8) that we expect in branch points. In particular, it shows high percentages of -2 `U` and the -3 `C` and is about 20-25 nucleotides from the 3′ splice site. This is encouraging but not definitive evidence of accuracy, which can only be obtained by experimentally mapping the branch points of some of these predictions.

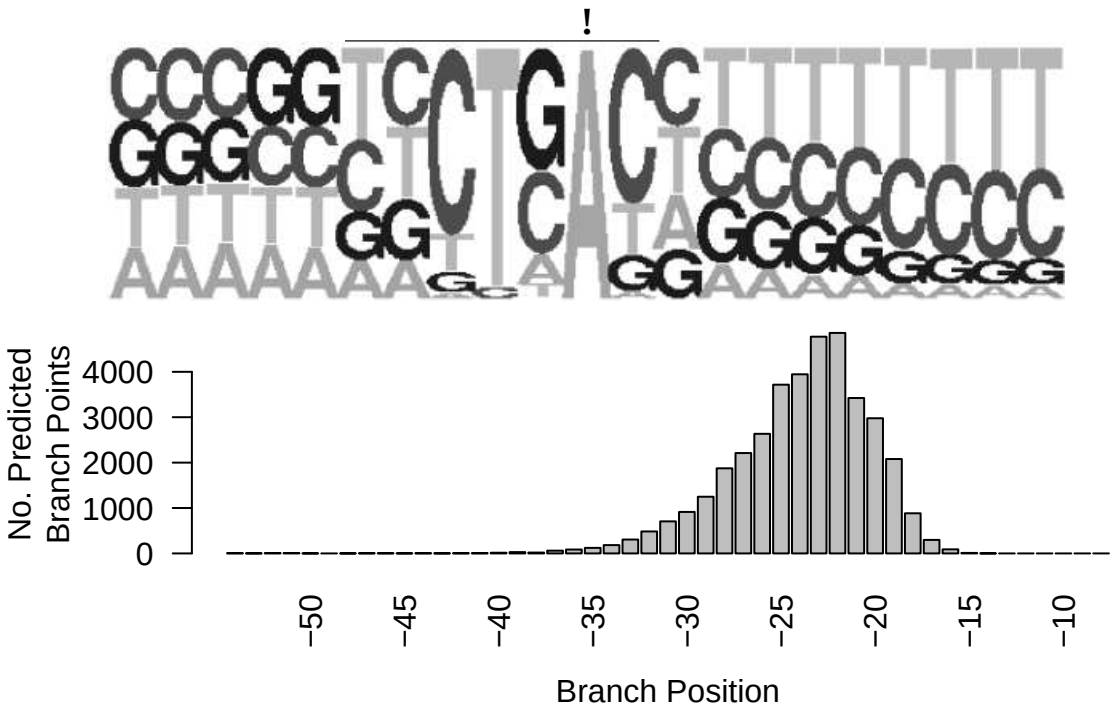


Figure 4-8: 38,000 novel human branch points. Top, pictogram of the BPS (exclamation point) and 10 flanking bases upstream and downstream for high-confidence predicted human branch points of SpliceGraph [57] introns. Overline indicates canonical branch site recognition positions relative to the predicted branch site (yeast consensus `TACTAAC`). Bottom, branch position relative to the 3′ splice site for same data set.

4.2.4 Branch Points and Exon Skipping

We next used SpliceGraph to determine which of our high-confidence branch points were associated with alternative exons. We formed a set of skipped exons (SE) by finding exons for which both isoforms (skipped and included) are supported by ESTs ($n = 3098$, or 5.1% of high-confidence branch points), and also a set of constitutive exons (CE) that are supported by at least two different ESTs and have no evidence of alternative splicing ($n = 12129$, 37.0%). Requiring support from two different ESTs for constitutive exons both increases the quality of the dataset and controls for expression bias: transcripts supported by more ESTs tend to be expressed at higher levels.

Skipped Exons Have Weaker Intron 3′ Ends than Constitutive Exons

We compared these two sets by several measures to see if there were any global differences between the branch points of skipped and constitutive exons. They were significantly different by several measures (see also Figures 4-9 and 4-10):

- BPSs of CE are more distant from the 3′ splice site than those of SE.
- CE have stronger branch point sequences when scored by the BPS state of BPA2.3.
- CE have stronger 3′ splice site when scored by IHMM20 or MaxEnt.
- CE have stronger branch sites when scored by highest posterior probability.

Although the magnitude of some of these differences is modest, it should be emphasized that these differences may be diminished by noise due to misprediction of branch sites or misannotation of SE/CE, and the true effect may be stronger. For example, the change in IHMM20 score is less than one bit, but it has been clearly documented in experimental systems that the most common result of weakening a splice site is exon skipping. Thus, in many ways the 3′ ends of skipped exons are weaker than those of constitutive exons.

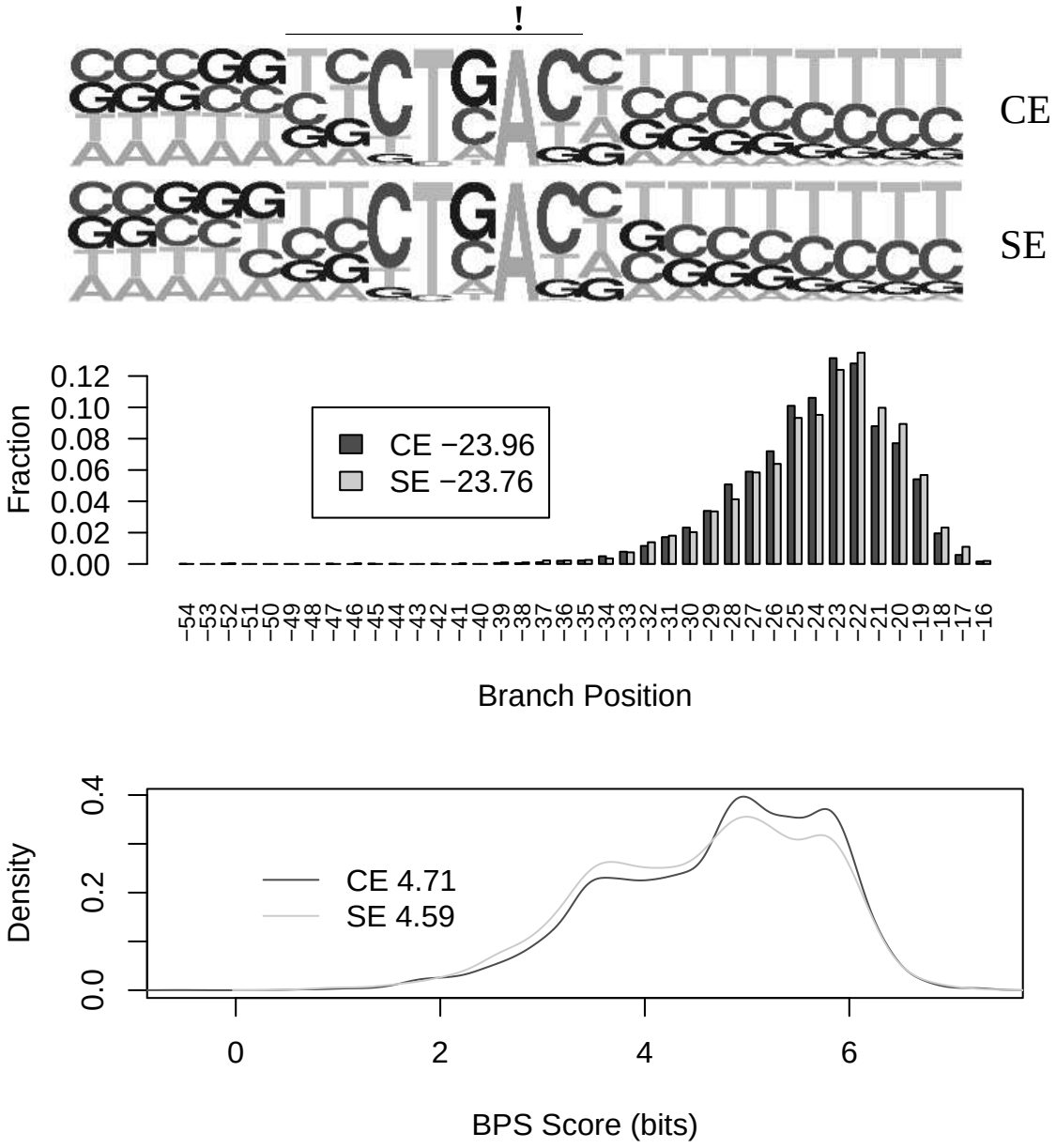


Figure 4-9: Comparison of branch sites of 3,098 skipped exons with 12,129 constitutive exons. Top, pictograms are of the -10 to +10 region relative to the predicted branch point (exclamation point), with the canonical branch site region overlined. Middle, CE have slightly longer branch distances than SE ($P = 0.0063$, means noted). Bottom, SE have weaker branch points than CE ($P = 2.3 \times 10^{-6}$, means noted).

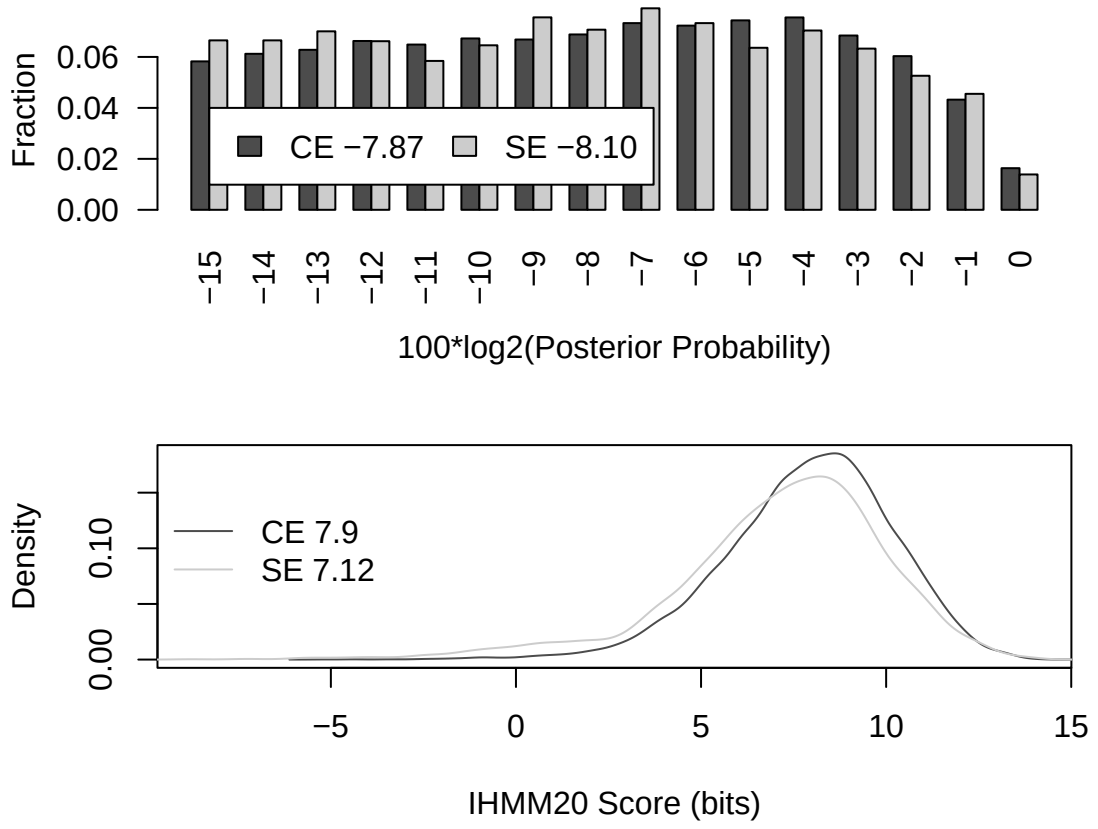


Figure 4-10: Comparison of branch sites of CE and SE, continued. Above, posterior probabilities of best branch positions. CE have stronger best branch positions ($P = 0.0077$, means noted). Below, IHMM20 score on -20 .. +3 region. CE have stronger PPT and 3′ splice sites ($P = 2.3 \times 10^{-42}$, means noted). The result with the MaxEnt model is similar.

SE Exhibit Improperly Ordered Elements

Continuing to compare skipped and constitutive exons, we next considered the difference in their sequence contents. We examined regions upstream and downstream of the predicted branch points: -45 to -6, termed `up40`, and +2 to +19, termed `down18`. We further fine-tuned `down18` in two ways. First, since we already know that SE have weaker PPT than CE, we created balanced subsets (Section 4.4.3) of SE and CE having the same IHMM20 score. These sets are about 10% smaller than the original sets (Figure 4-11, top) but now have essentially identical score distributions (Figure 4-11, bottom; compare with Figure 4-10, bottom). The reason for creating these sets is that we already know that CE have stronger $3'$ splice sites than SE, but we want to discover other differences. Next, we discarded those regions that overlapped the splice site `AG`, so that `down18` would only contain regions properly between the branch sequence and `AG` (Figure 4-11, `down18`).

For `up40` and `down18` and for $k = 1, \ldots, 6$, we performed χ^2 tests to identify k-mers enriched in SE or CE. Figure 4-12 shows those k-mers that showed a significant difference in their frequencies in sequence flanking predicted branch points of skipped versus constitutive exons.

The general trend in `up40` is that `A`+`T`-rich oligonucleotides are less frequent upstream of BPS of skipped exons than constitutive exons, and `G`+`C`-rich oligonucleotides are more frequent. In fact, the SE regions tend to be slightly more `G`+`C`-rich than the CE regions in general. The other group of k-mers we see are pyrimidine rich, for example `TTTCTC` and `TCTCCG`. These are enriched in the skipped exons relative to the constitutive exons. This suggests the possibility that pyrimidine tracts *upstream* of branch sites function as ISSs. In some ways this is analogous to the discovery that decoy $5'$ splice sites can suppress splicing when placed within exons [61], perhaps by competing with the real $5'$ splice sites for U1 binding. In a similar fashion, BPS-upstream pyrimidine tracts might function by competing with the BPS-downstream tracts for binding of U2AF65.

In `down18` the trend of `A`+`T`-rich oligonucleotide depletion in in skipped exons

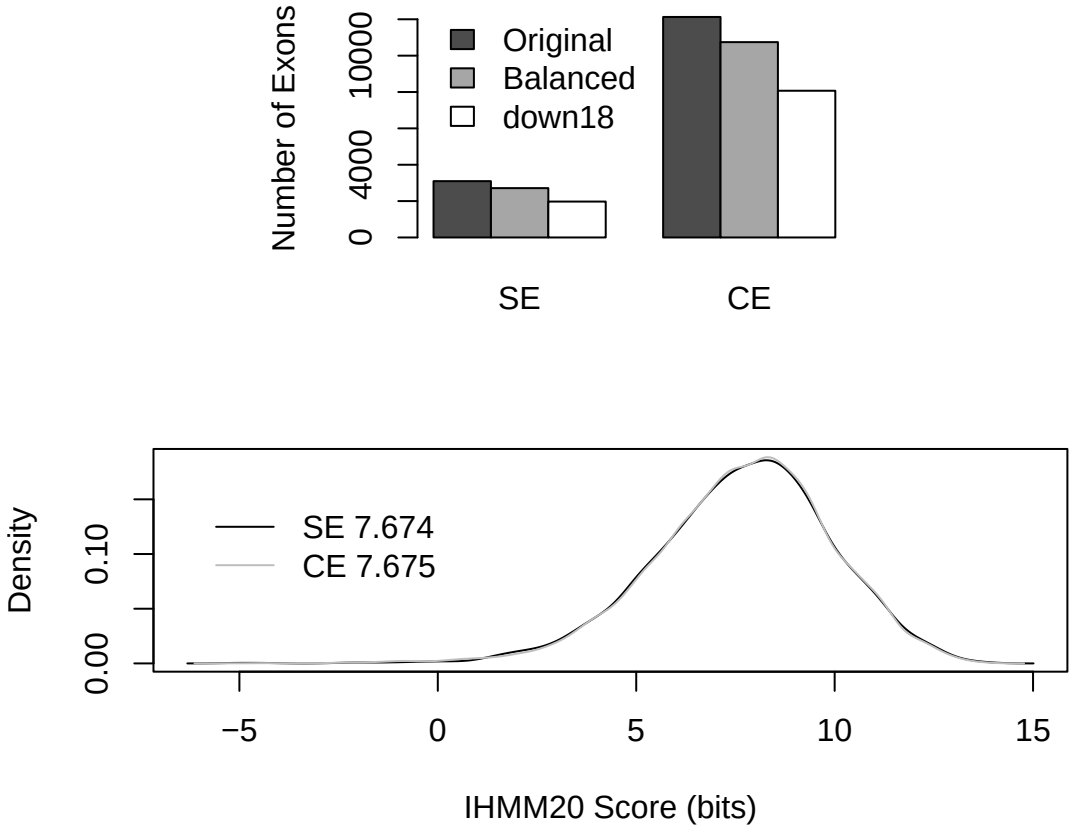


Figure 4-11: Score Balancing. Top, sizes of original, score-balanced, and `down18` sets. Bottom, IHMM20 score distributions of balanced sets (compare with Figure 4-10, bottom).

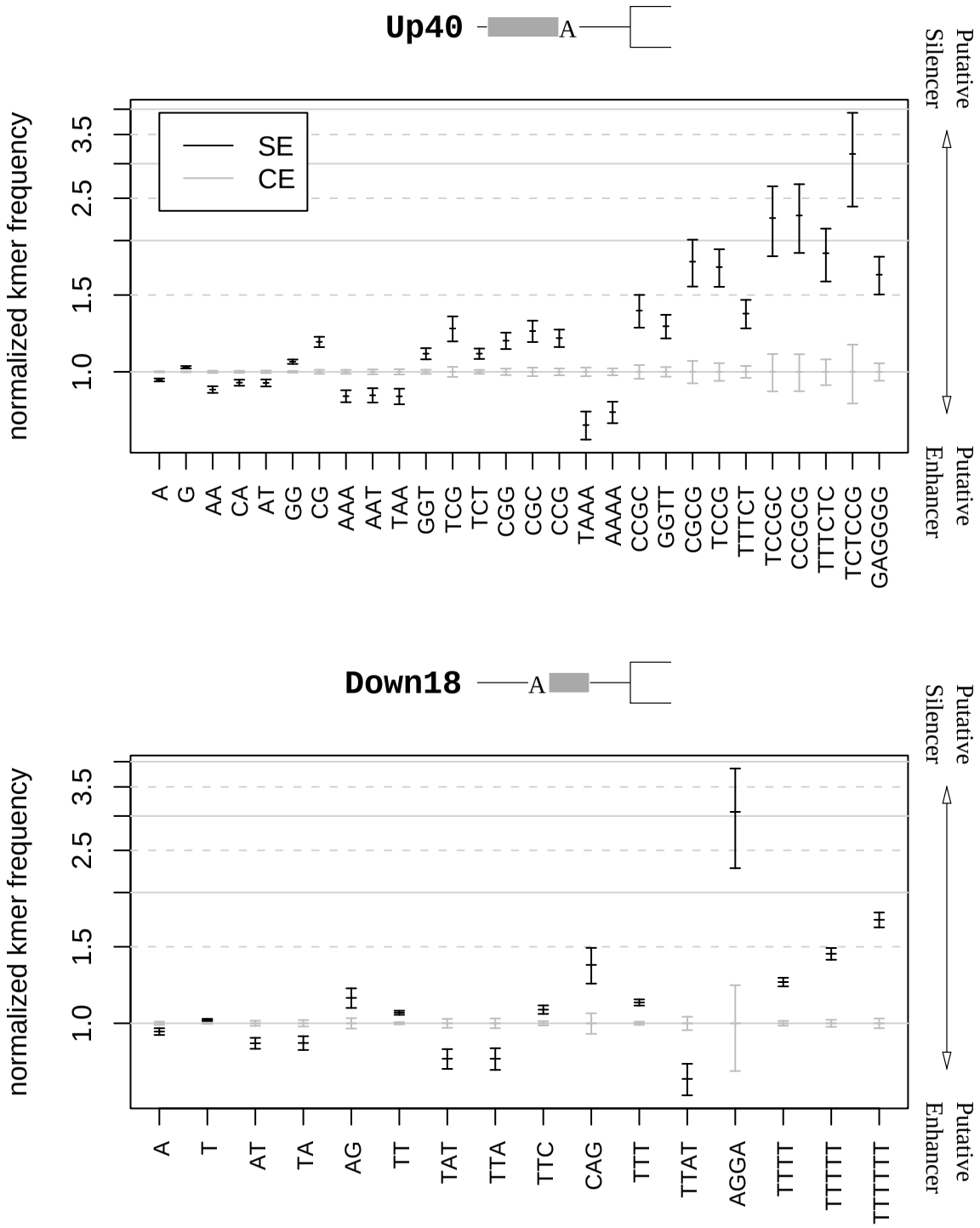


Figure 4-12: Enrichment of k-mers near BPS of constitutive and skipped exons. Top, `up40`. Bottom, `down18`. All frequencies were normalized to constitutive exon frequency for comparison. Error bars show binomial standard deviation.

persists (e.g. TAT and TTAT depletion in SE versus CE) but is weaker. Another group we see are perfect stretches of poly-T ($k = 3, 4, 5, 6$). These are enriched in SE versus CE, and increasingly enriched for larger k, suggesting that they function to suppress exon inclusion. U2AF65 has been previously shown to have a very high binding affinity for poly-T oligonucleotides [51], but our results suggest that this interaction actually suppresses the usage of the associated 3′ splice site. Finally, the last group are the AG-containing sequences such as CAG and AGGA. These may bind U2AF35 and compete with the correct 3′ splice site. These motifs are not enriched in up40, suggesting that AG alone is not efficient at binding U2AF35 in the absence of an upstream pyrimidine tract and/or branch point.

PPTs Occur Upstream of BPS in SE

Following the observation that pyrimidine-rich k-mers are enriched in SE relative to CE in up40, we asked whether skipped exons had a higher frequency of polypyrimidine tracts upstream of their BPS. To answer this question we created a linear sequence model consisting of positions -20 to -5 of IHMM20 flanked on both sides by the intronic state, which is the same in all models. We called this model IPI for the states it contains (Figure 4-13A). Positions -20 to -5 were chosen so as to exclude base -4, which does not generally have a pyrimidine composition, as well as subsequent bases.

We then defined a putative PPT-containing region as one with a positive maximum likelihood score. The median IPI score for the last 40 nucleotides of SE was 0.02 bits, so this is roughly the same. At first it may be surprising that the median score for this positive control set would be so close to zero, but it makes sense when one considers that the maximum likelihood score for IPI is a maximum over about 20 different parses (corresponding to possible starting positions for the putative PPT). Thus the probability of the maximum likelihood parse is in a sense penalized by $\log_2 20 \approx 4$ bits. If we add to this 4 bits another 4 bits for the conserved AG and another bit or two for the -3 and exonic positions then we have a typical 3′ splice site score of 9-10 bits. So the cutoff of 0 actually makes a lot of sense in the context of what we know about 3′ splice sites.

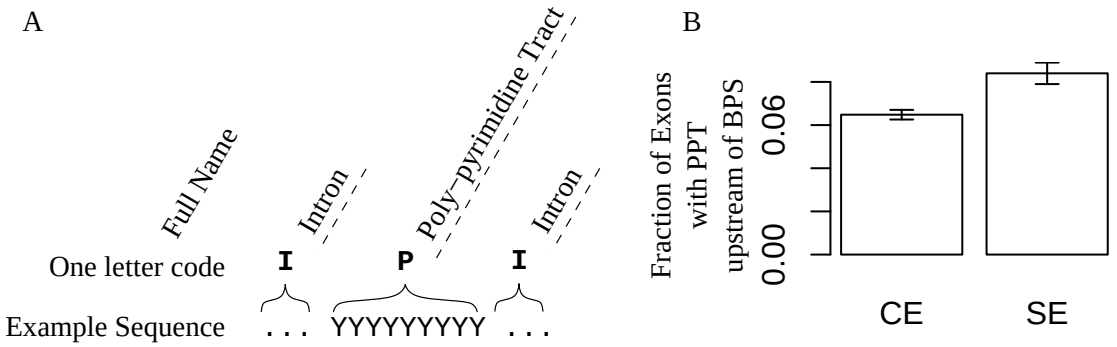


Figure 4-13: SE have PPT upstream of BPS. A, IPI linear sequence model for PPT within intronic sequence. IPI is made of positions -20 to -5 of IHMM20 flanked by the intronic state (which is the same for all models). The intronic states were both given uniform length distributions, so the score of a region is the LOD score of the highest scoring putative PPT in the region. B, For CE ($n = 12129$) and SE ($n = 3098$) sets, the `up40` region was scored with the IPI model using the Viterbi algorithm for LSMs, which gives the LOD score of the maximum likelihood parse. The fraction of introns with a positive score (hence containing a putative PPT) is shown, with binomial standard deviation error bars.

Having chosen this cutoff, we then scored the `up40` region of SE ($n = 12129$) and CE ($n = 3098$). 6.8% of CE and 8.9% of SE had putative PPT in `up40` so SE have significantly more putative PPT upstream of their branch points ($P = 7.0 \times 10^{-5}$, by χ^2 test, Figure 4-13B), in agreement with our interpretation of the k-mer enrichment data.

4.2.5 A Murine Branch Point Model Finds 32,000 Branch Points

We created a murine branch point model in order to be able to ask evolutionary questions about branch points, and also to be able to use comparative genomics to identify conserved features of intron $3'$ ends. To construct this model we took the human BPA2.3 and trained it for two iterations on a random set of about 10,000 murine $3'$ splice sites taken from the mouse (mm6) version of SpliceGraph. We then used this model, which we call mBPA, to predict branch points on all murine $3'$ splice sites in SpliceGraph, by taking those sites with a posterior probability at least 90%. This yielded 32,055 putative murine branch points.

In order to compare these murine branch points to human branch points we used

the liftOver utility available on the UCSC website, which can convert a range of coordinates in one genome to those on another. We took all BPS-3′ splice site ranges from the high-confidence mouse set and mapped them back to the human genome. Of 32,055 murine 3′ splice sites with high-confidence branch points, 3756 (11.7%) mapped onto human 3′ splice sites with high-confidence branch points. Of these, the predicted branch points of 3185 (84.8%) were at the same position in human and mouse. Thus it seems that branch point positions are highly conserved between human and mouse.

To address the question of whether branch sites are more conserved for SE or CE, we performed two χ^2 tests for homogeneity. First we examined whether the set of 3185 conserved BPS-3′ splice site ranges we found are enriched for SE or CE. 248 (8%) of SE were in this set, and 1130 (9.3%) of CE were in this set, a difference which is significant only at the 2.5% level. This is a surprise given the report of Kol et al finding branch sites of SE more conserved than those of CE [29]. However, those authors were examining sequence conservation specifically. If we take these 248 SE and 1130 CE whose BPS-3′ splice site ranges are conserved and ask what fraction of their sequences are conserved, we see that indeed the SE are more highly conserved at the sequence level (71 or 28.6% having perfectly conserved heptamer branch site) than CE (243 or 21.5% having perfectly conserved heptamer branch site), concurring with the other study. Thus we see that the extra branch site conservation in SE is specific to the sequence and does not apply to the position.

4.3 Discussion

We present here a new model for the 3′ ends of introns containing the branch site, the polypyrimidine tract, and the nucleotides near the 3′ splice site. Compared to previous models, this model is less dependent on very small training sets, and also provides a posterior probability indicating the confidence of each predicted branch site. Such a model should be used in any study of intronic regions near the 3′ splice site, as the sequences upstream and downstream of the branch point are vastly different.

4.3.1 Single-Adenosine Introns as a Branch Point Training Set

Most of the parameters of our BPA model were based on educated guesses based on our experience looking at authentic intron 3′ ends. These parameters were refined during three iterations of the EM training algorithm, so we reasoned that it was not essential that they be exactly correct. However, for the initial guesses of branch point emission probabilities we generated a set of introns containing exactly one adenosine in the branch point permissible region (intron positions +40..-10). The consensus sequence at positions -5..+1 relative to the branch point adenosine strongly resembles previously reported branch point sequences, except for an artefactual avoidance of adenosine at the other positions. As such, we are highly confident that these represent a set of correct, albeit biased branch sites. This set can be used by other researchers interested in studying high-confidence branch points apart from those already published in the literature mapped by biochemical or genetic methods.

4.3.2 Model Validation

BPA2.3 can achieve the same or better Exact Accuracy as the MaxEnt model [64] when used as part of the Exonscan splicing simulator [61] to predict exons in a test set of about 2000 constitutively spliced transcripts. Furthermore, BPA2.3 correctly predicts 15 of 25 experimentally mapped branch point sites mined from the literature when the maximum posterior probability position is chosen. Of the 25 branch point sites, 5 had a position with 90% or greater posterior probability, and all of the 5 were correctly predicted (Figure 4-7). Therefore we chose a cutoff of 90% posterior probability for our high-confidence set of branch points.

4.3.3 Prediction of High-Confidence Branch Points for 40,000 Authentic Human $3'$ Splice Sites

Out of roughly 275,000 human $3'$ splice sites in the SpliceGraph database [57], 40,000 exhibited high-confidence branch points as scored by BPA2.3. This set showed expected sequence composition flanking the branch sites as well as expected distribution of distance to the $3'$ splice site (Figure 4-8). A similar model specialized for mouse introns predicted roughly 30,000 high-confidence branch points upstream of $3'$ splice sites in that organism. Of these splice sites, approximately 4000 (12%) aligned to human splice sites with high-confidence predicted branch sites, and these branch sites were also aligned for approximately 85% of the aligned $3'$ splice sites. These data point to a high conservation of branch sites in evolution between human and mouse, suggesting that mutations in the branch site are often deleterious and not easily compensated by activating latent branch points.

4.3.4 Skipped Exons Exhibit Misordered Sequence Elements

As an application of our BPA2.3 model we studied the branch sites upstream of exons in SpliceGraph associated with either alternatively skipped (SE, roughly 3,000) or constitutively included (CE, roughly 12,000) exons. As has been previously observed, the skipped exons had weaker $3'$ splice sites when scored by either MaxEnt or our second order inhomogeneous Markov model ($P = 2.3 \times 10^{-42}$). In addition, the branch sites of SE had lower sequence scores ($P = 2.3 \times 10^{-6}$) and were closer to the $3'$ splice sites ($P = 0.0063$) than constitutive exons. Finally, the entire region scored by BPA2.3 had a lower score in SE than CE when scored by maximum posterior probability ($P = 0.0077$). These results extend previous observations about the weakness of splice sites of skipped exons relative to constitutive exons.

To see if we could identify specific motifs related to exon skipping or inclusion, we looked at k-mer frequencies appearing in upstream and downstream windows of branch sites of constitutive exons (CE) and skipped exons (SE). Interestingly, we found pyrimidine-rich sequences upstream of branch sites enriched in SE relative to

CE, and sequences containing the AG dinucleotide between branch sites and 3′ splice sites enriched in SE relative to CE. Typically we expect the region between branch sites and 3′ splice sites to contain polypyrimidine tracts and to lack AG dinucleotides except at the last two positions of the intron. Thus a salient feature of SE relative to CE is the misordering of 3′-end elements. To further test this interpretation we looked specifically for high-scoring polypyrimidine tracts using our inhomogeneous Markov model and found that they are enriched upstream of branch points of SE relative to CE ($P = 7.0 \times 10^{-5}$), in corroboration of our interpretation.

Finally, we examined the conservation of branch sites of SE and CE between human and mouse. As previously reported, the heptamer branch site sequences of SE are more highly conserved than those of CE. However, the positions of those branch sites in sequence alignments relative to the 3′ splice site are conserved at an approximately equal rate in SE and CE, indicating that the geometry of branch sites is highly conserved and therefore highlighting the importance of branch sites in evolution.

The BPA model is thus a high-confidence method to predict branch points at the 3′ end of introns. It could be used in medical applications to predict branch sites in disease-associated mutations. It will be interesting to continue studying the evolution of branch points using this model. Finally, since the regions upstream and downstream of branch sites are drastically different, such a model will be important in the analysis of the 3′ ends of introns flanking alternatively spliced exons.

4.4 Methods

4.4.1 The Theory of Linear Sequence Models

Definition

We develop here the theory for a class of probability distributions which we call *linear sequence models* (LSMs)*. As we shall see, there are many analogies between

*These are actually a special case of a class of distributions known as hidden semi-Markov models

the HMMs of Section 1.6.3 and LSMs. In LSMs we have S states and a finite alphabet $\mathcal{A}$. In general, each state is a probability distribution P_s on all finite words taken from $\mathcal{A}$: $\sum_{l\geq 0}\sum_{x\in\mathcal{A}^l} P_s(x) = 1$. In our case, we specify state s by a Markov model of some order k_s, specified by transition matrix $T_s \in \mathbb{R}^{\mathcal{A}^{k_s}\times\mathcal{A}}$, and, independently, a length distribution L_s ($\sum_{l\geq 0} L_s(l) = 1$). Therefore for a word $x \in \mathcal{A}^l$ of length l, $P_s(x) = L_s(l)M_s(x)$, where $M_s(x) = P_s(x|l)$ denotes the Markov probability $\prod_{i=1}^{l} T_s(\underline{x_{i-k_s}\cdots x_{i-1}} \to x_i)$.

Given the L_s and M_s, the model generates sequence by selecting lengths $l_1, \ldots, l_S \geq 0$ each according to $L_s(l_s)$, generating subsequences x_s of those lengths according to the Markov probabilities M_s, and concatenating those subsequences.

Sequence Parsing

Given the sequence $\underline{\sigma}$ of length n the *forward algorithm for LSMs* fills the matrix F, which represents the probability of an initial segment of states emitting an initial segment of sequence:

$$\begin{aligned} f_{is} &= P(\text{states } 1\ldots s \text{ emit } \underline{\sigma_1\cdots\sigma_i}) \\ &= \sum_{0=i_0\leq\cdots\leq i_s=i} \left[\prod_{1\leq t\leq s} L_t(i_t - i_{t-1}) M_t(\underline{\sigma_{i_{t-1}+1}\cdots\sigma_{i_t}})\right]. \end{aligned} \tag{4.1}$$

Here t iterates over the states from 1 to s, and (i_t) represent the boundaries of the states, so that state t emits the sequence $\underline{\sigma_{i_{t-1}+1}\cdots\sigma_{i_t}}$, which is of length $i_t - i_{t-1}$. In English, the second line of Equation 4.1 is the sum over all possible partitions of sequence $\underline{\sigma_1\cdots\sigma_i}$ among the first s states of the product over those states of emission probabilities for the corresponding subsequences. This sum can be quickly calculated by recursing over the length of the last state:

$$\begin{aligned} f_{is} &= \sum_{l=0}^{i} f_{i-l,s-1}\ L_s(l)\ M_s(\underline{\sigma_{i-l+1}\cdots\sigma_i}) \\ f_{i0} &= \delta(i) \\ f_{0s} &= \delta(s). \end{aligned} \tag{4.2}$$

where $\delta(\cdot)$ is the Kronecker Delta function, valued 1 at 0 and 0 elsewhere. Just as with simple HMMs, f_{nS} gives the total probability of the sequence $\underline{\sigma}$. Similarly, the *backward algorithm for LSMs* fills the matrix B where $b_{is} = P(\text{states } s \dots S \text{ emit } \underline{\sigma_i \cdots \sigma_n})$. The forward and backward algorithm run in $O(n^2 S)$ time and require $O(nS)$ space, compared to $O(nS^2)$ time and $O(nS)$ space for simple HMMs.

The *Viterbi algorithm for LSMs* gives the maximum log likelihood parse of $\underline{\sigma}$. For LSMs, a parse is specified by the boundaries of the states. The Viterbi algorithm fills the matrix V defined by

$$v_{is} = \max_{0=i_0 \le \cdots \le i_s = i} \left[\sum_{1 \le t \le s} \log L_t(i_t - i_{t-1}) + \log M_t(\underline{\sigma_{i_{t-1}+1} \cdots \sigma_{i_t}}) \right].$$

Note the similarity with Equation 4.1. The mapping $(\sum, \prod) \to (\max, \sum)$ is known as tropicalization[†] ([42]). V is filled by tropicalizing Equations 4.2:

$$\begin{aligned} v_{is} &= \max_{l=0}^{i} \left[v_{i-l,s-1} + \log L_s(l) + \log M_s(\underline{\sigma_{i-l+1} \cdots \sigma_i}) \right] \\ v_{i0} &= \log \delta(i) \\ v_{0s} &= \log \delta(s), \end{aligned}$$

where $\log \delta(\cdot)$ is valued 0 at 0 and $-\infty$ elsewhere. The viterbi algorithm has the same complexity as the forward and backward algorithms, namely $O(n^2 S)$ time and $O(nS)$ space.

The actual maximum likelihood parse can be reconstructed by remembering the lengths that achieve the maxima. This requires an extra $O(nS)$ space, which is not problematic, so does not necessitate the traceback algorithm sometimes used in stochastic context free grammars [24].

F and B can be used to calculate *posterior probabilities* for a state s given the sequence $\underline{\sigma}$. Thus for $i \le j+1$ the posterior probability of the state s emitting $\underline{\sigma_i \cdots \sigma_j}$

[†]since it was first studied by the Brazilian mathematician Imre Simon

is given by

$$P_s^{\text{post}}(i,j) = f_{i-1,s-1}\; L_s(j-i+1)\; M_s(\underline{\sigma_i \cdots \sigma_j})\; b_{j+1,s+1}. \tag{4.3}$$

This expression is very useful for evaluating different positions in our sequence as potential branch points, and also important in the training algorithm.

Training

Training is accomplished by the *Baum-Welch algorithm for LSMs*. Given sequences $\underline{\sigma}^{(i)}$ for $i = 1 \ldots N$ we have to determine the length distribution $\widehat{L_s}$ and Markov probabilities $\widehat{T}_s$ for each state s. Since, for each sequence, $P_s^{\text{post}}(i,j)$ (Equation 4.3) is a probability distribution on the intervals $i \leq j+1$, it also induces a probability distribution on the lengths, namely $L_s^{\text{post}}(l;\underline{\sigma}) = \sum_{i-j+1=l} P_s^{\text{post}}(i,j;\underline{\sigma})$, where $P_s^{\text{post}}(i,j;\underline{\sigma})$ denotes the posterior probability of the state s on the interval $[i,j]$ given the sequence $\underline{\sigma}$.

The trained length distribution can therefore be calculated by averaging the posterior length distributions of all the sequences:

$$\widehat{L_s}(l) = \frac{1}{N} \sum_{i=1}^{N} L_s^{\text{post}}(l;\underline{\sigma}^{(i)}).$$

The trained Markov transition probabilities are calculated in a similar fashion. For $x \in \mathcal{A}^{k_s}$ and $y \in \mathcal{A}$, we would like to set the transition probability of $\underline{x}$ to y, $\widehat{T}_s(\underline{x} \to y)$, to be the ratio of expected counts of $\underline{xy}$ and $\underline{x}$ in state s. We can calculate these expected counts by weighing the actual occurrences of these words by posterior probabilities of the intervals containing them, being careful about which positions are conditional and which are emitted by the Markov model. Letting $\#\underline{z}(\underline{\tau})$ denote the number of times the word $\underline{z}$ appears among the letters in the sequence $\underline{\tau}$, the expected

counts of $\underline{xy}$ and $\underline{x}$ in state s in the sequence $\underline{\sigma}$ are given respectively by

$$\widehat{N_s}(\underline{x} \to y; \underline{\sigma}) = \sum_{i \le j+1} P_s^{\text{post}}(i, j; \underline{\sigma})\ \#\underline{xy}(\underline{\sigma_{i-k_s} \cdots \sigma_j}) \quad \text{and}$$
$$\widehat{N_s}(\underline{x} \to \cdot; \underline{\sigma}) = \sum_{i \le j+1} P_s^{\text{post}}(i, j; \underline{\sigma})\ \#\underline{x}(\underline{\sigma_{i-k_s} \cdots \sigma_{j-1}}).$$

Finally we can calculate the trained Markov transition probabilities by

$$\widehat{T_s}(\underline{x} \to y) = \frac{\sum_{i=1}^{N} \widehat{N_s}(\underline{x} \to y; \underline{\sigma}^{(i)})}{\sum_{i=1}^{N} \widehat{N_s}(\underline{x} \to \cdot; \underline{\sigma}^{(i)})}.$$

With these new $\widehat{T_s}$ and $\widehat{L_s}$ another iteration of training can be performed if desired.

4.4.2 The BPA implementation

We actually use a *conditional* version of the Linear Sequence Model algorithms, which means that all probabilities are conditioned on the first few nucleotides of the sequence. This distinction is probably only important in the implementation of the model and not for any of the actual results.

In practice LSMs are used in three different predictive modes. The first is posterior probability which is described above. The second is total probability, in which we use the forward probability. To get a LOD score for the total probability we use the Markov model from the intronic state as our background distribution, matching the length distributions. This means that we need to generate the length distribution of the entire LSM, but that can be quickly done by convoluting the length distributions of the individual states according to the recursion

$$\mathscr{L}_s(l) = \sum_{l'} \mathscr{L}_{s-1}(l - l') L_s(l')$$
$$\mathscr{L}_1(l) = L_1(l),$$

where $\mathscr{L}_s(l)$ represents the probability that the length of sequence emitted by the first s states is l; $L_s(\cdot)$ is the length distribution of state s, as in the previous section;

and l' ranges from the minimum to maximum lengths of the state. Therefore, the final LOD score for a sequence $\underline{\sigma}$ of length n would be given by

$$\log f_{nS} - \log \mathscr{L}_S(n) - \log M_I(\underline{\sigma}), \tag{4.4}$$

where f_{nS} is the total (forward) probability of the $\underline{\sigma}$ for that LSM, and $M_I(\cdot)$ denotes the Markov probability using the Markov model from the intronic state.

Finally, in the case where we would like a score for the maximum likelihood parse, we use Equation 4.4, replacing $\log f_{nS}$ with v_{nS}, the log likelihood of that parse given by the Viterbi algorithm for LSMs.

4.4.3 Balanced Datasets by Relative Reduction

Suppose we have two datasets X and Y for which each element contains a number of properties, one of which v is an integer-valued statistic that we'd like to control, letting $v(z)$ represent the value of this property for an element $z \in X$ or $z \in Y$. One example of this problem is when X and Y are $3'$ splice sites of skipped and constitutive exons, and v is their splice site score.

We can create balanced subsets of X and Y having the same distribution of v using the following stochastic algorithm. First define the smoothed density function f_Z of v ($Z = X$ or Y) by fixing a window radius w and letting

$$f_Z(z) = \frac{|\{z' \in Z : z - w \leq z' \leq z + w\}|}{|Z|}.$$

For the $3'$ splice site case we used $w = 0.2$ bits. Next define the *relative reduction functions* for the two sets, $r_X(\cdot)$ and $r_Y(\cdot)$. This quantity represents, for a given value v, the factor by which we need to reduce the density of that set in order to match the density of the other set, or 1 if the density is already less than the other set:

$$r_X(v) = \begin{cases} \frac{f_Y(v)}{f_X(v)} & \text{if } f_X(v) > f_Y(v) \\ 1 & \text{else} \end{cases},$$

and $r_Y(v)$ is defined analogously.

With the relative reductions calculated, we then form the subsets by taking each element $x \in X$ with probability $r_X(x)$ and likewise for Y.

Bibliography

[1] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.

[2] M. L. Berbee and J. W. Taylor. *The Mycota, Volume VII: Systematics and evolution, part B.*, chapter Fungal Molecular Evolution: Gene Trees and Geologic Time. New York: Springer-Verlag, 2000.

[3] S. M. Berget, C. Moore, and P. A. Sharp. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A*, 74(8):3171–5, 1977.

[4] J. D. Boeke, D. J. Garfinkel, C. A. Styles, and G. R. Fink. Ty elements transpose through an RNA intermediate. *Cell*, 40(3):491–500, 1985.

[5] C. B. Burge, R. A. Padgett, and P. A. Sharp. Evolutionary fates and origins of U12-type introns. *Mol Cell*, 2(6):773–85, 1998.

[6] C. B. Burge, T. Tuschl, and P. A. Sharp. *The RNA World*, chapter Splicing of precursors to mRNA by the spliceosome, pages 525–560. Cold Spring Harbor Press, 1999.

[7] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*, 31(13):3497–500, 2003.

[8] L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, 12(1):1–8, 1977.

[9] J. Q. Clement, L. Qian, N. Kaplinsky, and M. F. Wilkinson. The stability and fate of a spliced intron from vertebrate cells. *RNA*, 5(2):206–20, 1999.

[10] A. Coghlan and K. H. Wolfe. Origins of recently gained introns in Caenorhabditis. *Proc Natl Acad Sci U S A*, 101(31):11362–7, 2004.

[11] L.R. Coulter, M.A. Landree, and T.A. Cooper. Identification of a new class of exonic splicing enhancers by in vivo selection. *Mol Cell Biol*, 17(4):2143–50, 1997.

[12] L. Dollo. Les lois de l'evolution. *Bull Soc Belge Geol Pal Hydr*, 7:164–166, 1893.

[13] D. Dressler and H. Potter. *Discovering Enzymes*. Scientific American Library Series, 1991.

[14] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge UP, 1998.

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

[16] J.S. Farris. Phylogenetic analysis under Dollo's law. *Syst Zool*, 26:77–88, 1977.

[17] C. Federico, L. Andreozzi, S. Saccone, and G. Bernardi. Gene density in the Giemsa bands of human chromosomes. *Chromosome Res*, 8(8):737–46, 2000.

[18] A. Fedorov, S. Roy, L. Fedorova, and W. Gilbert. Mystery of intron gain. *Genome Res*, 13(10):2236–41, 2003.

[19] G. R. Fink. Pseudogenes in yeast? *Cell*, 49(1):5–6, 1987.

[20] B. R. Graveley and T. Maniatis. Arginine/Serine-Rich Domains of SR Proteins Can Function as Activators of Pre-mRNA Splicing. *Mol Cell*, 1(1):765–771, 1998.

[21] K. Hartmuth and A. Barta. Unusual branch point selection in processing of human growth hormone pre-mRNA. *Mol Cell Biol*, 8(5):2011–20, 1988.

[22] F. Hartung, F. R. Blattner, and H. Puchta. Intron gain and loss in the evolution of the conserved eukaryotic recombination machinery. *Nucleic Acids Res*, 30(23):5175–81, 2002.

[23] D. S. Heckman, D. M. Geiser, B. R. Eidell, R. L. Stauffer, N. L. Kardos, and S. B. Hedges. Molecular evidence for the early colonization of land by fungi and plants. *Science*, 293(5532):1129–33, 2001.

[24] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie*, 125:167–188, 1994.

[25] D. G. Hwang and P. Green. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A*, 101(39):13994–4001, 2004.

[26] C. Ibrahim el, T. D. Schaal, K. J. Hertel, R. Reed, and T. Maniatis. Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. *Proc Natl Acad Sci U S A*, 102(14):5002–7, 2005.

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

[28] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Res*, 12(6):996–1006, 2002.

[29] G. Kol, G. Lev-Maor, and G. Ast. Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. *Hum Mol Genet*, 14(11):1559–68, 2005.

[30] J. Kralovicova, S. Houngninou-Molango, A. Kramer, and I. Vorechovsky. Branch site haplotypes that control alternative splicing. *Hum Mol Genet*, 13(24):3189–202, 2004.

[31] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–14, 1993.

[32] A. Llopart, J. M. Comeron, F. G. Brunet, D. Lachaise, and M. Long. Intron presence-absence polymorphism in Drosophila driven by positive Darwinian selection. *Proc Natl Acad Sci U S A*, 99(12):8121–6, 2002.

[33] J. M. Logsdon, Jr, A. Stoltzfus, and W. F. Doolittle. Molecular evolution: recent cases of spliceosomal intron gain? *Curr Biol*, 8(16):R560–3, 1998.

[34] J. M. Logsdon, Jr, M. G. Tyshenko, C. Dixon, J. D-Jafari, V. K. Walker, and J. D. Palmer. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc Natl Acad Sci U S A*, 92(18):8507–11, 1995.

[35] T. Mourier and D. C. Jeffares. Eukaryotic intron loss. *Science*, 300(5624):1393, 2003.

[36] M. Nei and S. Kumar. *Molecular Evolution and Phylogenetics*. New York: Oxford University Press, 2000.

[37] A. K. Ng, T. M. Block, B. Aiamkitsumrit, M. Wang, E. Clementi, T. T. Wu, J. M. Taylor, and Y. H. Su. Construction of a herpes simplex virus type 1 mutant with only a three-nucleotide change in the branchpoint region of the latency-associated transcript (LAT) and the stability of its two-kilobase LAT intron. *J Virol*, 78(22):12097–106, 2004.

[38] C. B. Nielsen, B. Friedman, B. Birren, C. B. Burge, and J. E. Galagan. Patterns of intron gain and loss in fungi. *PLoS Biol*, 2(12):e422, 2004.

[39] J. C. Noble, C. Prives, and J. L. Manley. In vitro splicing of simian virus 40 early pre mRNA. *Nucleic Acids Res*, 14(3):1219–35, 1986.

[40] F. C. Oberstrass, S. D. Auweter, M. Erat, Y. Hargous, A. Henning, P. Wenter, L. Reymond, B. Amir-Ahmady, S. Pitsch, D. L. Black, and F. H. Allain. Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science*, 309(5743):2054–7, 2005.

[41] R. J. O'Neill, F. E. Brennan, M. L. Delbridge, R. H. Crozier, and J. A. Graves. De novo insertion of an intron into the mammalian sex determining gene, SRY. *Proc Natl Acad Sci U S A*, 95(4):1653–7, 1998.

[42] L. Pachter and B. Sturmfels, editors. *Algebraic Statistics for Computational Biology*. Cambridge UP, 2005.

[43] W. G. Qiu, N. Schisler, and A. Stoltzfus. The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol Biol Evol*, 21(7):1252–63, 2004.

[44] R. Reed. The organization of 3' splice-site sequences in mammalian introns. *Genes Dev*, 3(12B):2113–23, 1989.

[45] R. Reed. Personal Communication. 2005.

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

[47] H. M. Robertson. The large srh family of chemoreceptor genes in Caenorhabditis nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. *Genome Res*, 10(2):192–203, 2000.

[48] I. B. Rogozin, Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, and E. V. Koonin. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol*, 13(17):1512–7, 2003.

[49] S. W. Roy, A. Fedorov, and W. Gilbert. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci U S A*, 100(12):7158–62, 2003.

[50] D. Schmucker, J. C. Clemens, H. Shu, C. A. Worby, J. Xiao, M. Muda, J. E. Dixon, and S. L. Zipursky. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101(6):671–84, 2000.

[51] R. Singh, J. Valcarcel, and M. R. Green. Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science*, 268(5214):1173–6, 1995.

[52] C. W. Smith, T. T. Chu, and B. Nadal-Ginard. Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Mol Cell Biol*, 13(8):4939–52, 1993.

[53] C. W. Smith and B. Nadal-Ginard. Mutually exclusive splicing of alpha-tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing. *Cell*, 56(5):749–58, 1989.

[54] C. W. Smith, E. B. Porro, J. G. Patton, and B. Nadal-Ginard. Scanning from an independently specified branch point defines the 3' splice site of mammalian introns. *Nature*, 342(6247):243–7, 1989.

[55] J. Southby, C. Gooding, and C. W. Smith. Polypyrimidine tract binding protein functions as a repressor to regulate alternative splicing of alpha-actinin mutally exclusive exons. *Mol Cell Biol*, 19(4):2699–711, 1999.

[56] M. Spingola, L. Grate, D. Haussler, and M. Ares, Jr. Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. *RNA*, 5(2):221–34, 1999.

[57] M. Stadler and R. Sandberg. SpliceGraph, a database of transcript-confirmed gene isoforms. Unpublished Data.

[58] H. Suzuki, Y. Jin, H. Otani, K. Yasuda, and K. Inoue. Regulation of alternative splicing of alpha-actinin transcript by Bruno-like proteins. *Genes Cells*, 7:133–141, 2002.

[59] R. Tacke and J.L. Manley. The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *EMBO J*, 14(14):3540–51, 1995.

[60] T. N. Taylor, H. Hass, and H. Kerp. The oldest fossil ascomycetes. *Nature*, 399(6737):648, 1999.

[61] Z. Wang, M.E. Rolish, G. Yeo, V. Tung, M. Mawson, and C.B. Burge. Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–45, 2004.

[62] Z. Wang, X. Xiao, E. Van Nostrand, and C. B. Burge. General and specific functions of exonic splicing silencers in splicing control. *Mol Cell*, 23(1):61–70, 2006.

[63] J. C. Webb, D. D. Patel, C. C. Shoulders, B. L. Knight, and A. K. Soutar. Genetic variation at a splicing branch point in intron 9 of the low density lipoprotein (LDL)-receptor gene: a rare mutation that disrupts mRNA splicing in a patient with familial hypercholesterolaemia and a common polymorphism. *Hum Mol Genet*, 5(9):1325–31, 1996.

[64] G. Yeo and C. B. Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*, 11(2-3):377–94, 2004.

[65] G. Yeo, S. Hoon, B. Venkatesh, and C. B. Burge. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A*, 101(44):15700–5, 2004.

[66] X.H. Zhang and L.A. Chasin. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*, 18(11):1241–50, 2004.

Appendix A

Supplementary Figures

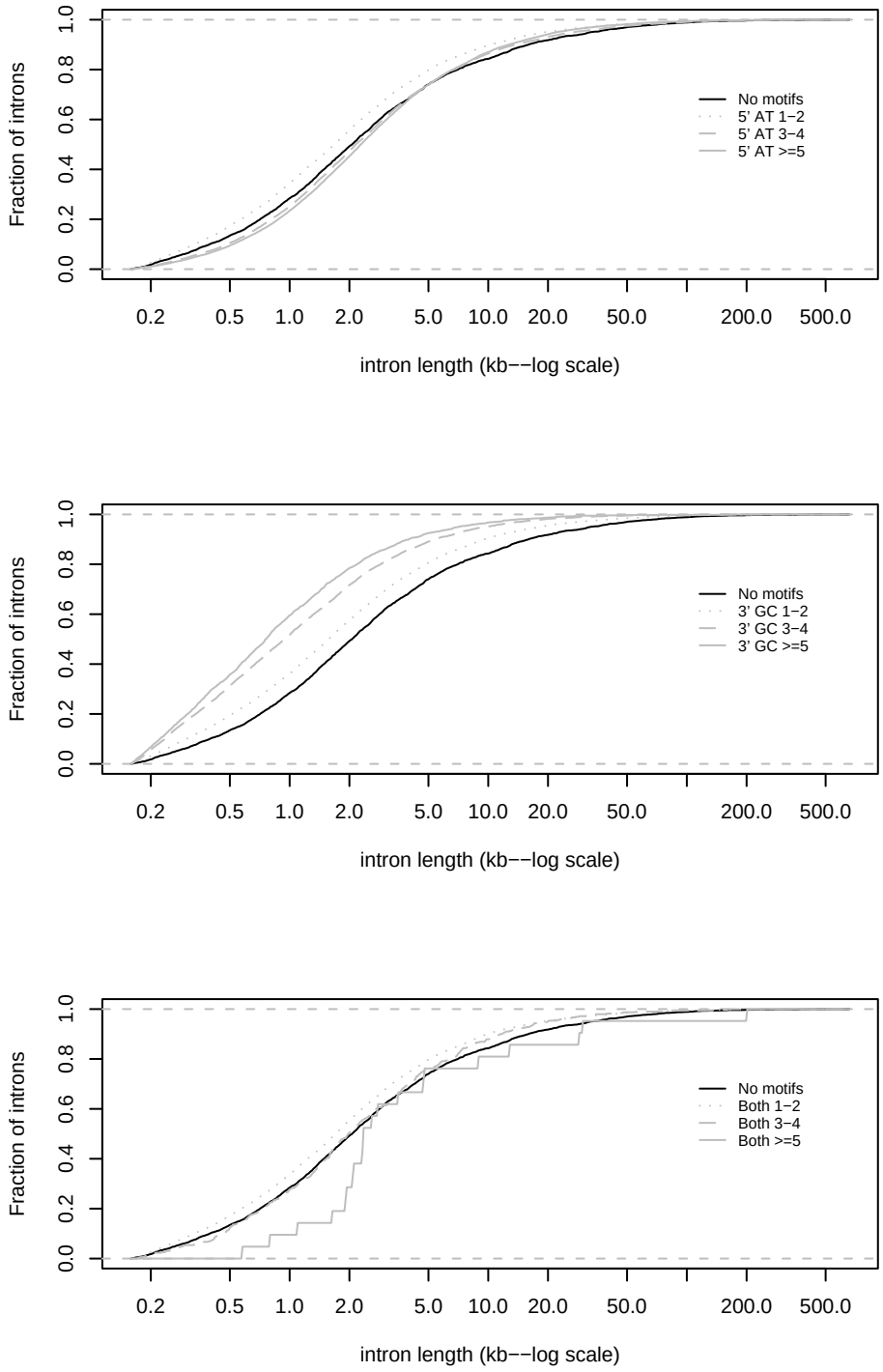


Figure A-1: Introns with many `AU`/`GC` motif pairs do not show the same tendency toward length as those with many `GC`/`AU` pairs (see Figure 3-11). ECDF of lengths of constituive intron with increasing numbers of `AT` motif near the 3′ splice site (top), `GC` motifs near the 3′ splice site (middle), or both motifs at their corresponding splice sites.