Education

Strategies for Identifying RNA Splicing Regulatory Motifs and Predicting Alternative Splicing Events

Dirk Holste*, Uwe Ohler*

A Tutorial in PLoS Computional Biology

Gene Expression and RNA Splicing

The regulation of gene expression is a ubiquitous phenomenon and is involved in virtually every process central to an organism, ranging from the fertilization of germ cells, across the cell cycle, to stimuli–response pathways or apoptosis. To control the expression of genes under such diverse contexts, regulation occurs on different cellular levels and involves a series of complex biochemical mechanisms that one can broadly classify into transcription, RNA processing and cytoplasmic transport, and post-transcriptional control and translation. While a series of distinct machineries is involved in controlling gene expression at each level, these complex circuits bear signs of interconnectedness [1].

In higher eukaryotes, splicing constitutes a critical mode for the regulation of gene expression at the level of RNA processing [2–4]. The large majority of eukaryotic protein-coding genes are transcribed as precursors of messenger RNAs (pre-mRNAs), in which exons are separated from each other by intervening regions of non-protein–coding information (introns), which have to be correctly spliced out to produce a mature mRNA. Splicing of pre-mRNAs occurs in a two-step reaction (Figure 1). In the first step, the message is cleaved at the 5′ end of an intron, and this 5′ end is linked to the branch point, which is typically in close proximity upstream of the 3′ end of the intron. In the second step, the mRNA intermediate is cleaved at the 3′ splice site (3′ss), exons are ligated, and the intron lariat is released [2]. During later stages of spliceosome assembly, the 5′ss and 3′ss pair and interact (typically across the exon, but pairing across an intron can occur), supported by general and specific splicing factors that recognize them. Typical mammalian genes span tens of thousands of nucleotides, with on average nine exons and protein-coding regions on the order of a thousand nucleotides, thus embedding "exon islands" within a large "sea" of noncoding nucleotides that have to be accurately recognized for correct splicing and exon ligation. This important task is executed in the nucleus by the spliceosome, a large ribonucleoprotein (RNP) complex that involves five small nuclear RNAs and potentially hundreds of proteins, the core components of which are highly conserved across metazoan genomes [5].

Signals that specify exon–intron junctions are located at the termini of introns. *cis*-Acting nucleic-acids elements are located at the 5′ss, branch point, and 3′ss, and guide the spliceosome. Almost all introns are characterized by /GT and AG/ termini at the 5’ss and 3’ss, respectively (U2-type introns). In addition to the canonical /GT and AG/ termini, between four and seven nucleotides (5′ss) and up to about 20 nucleotides (3′ss) typically contain information for splicing. A small fraction of U2-introns exhibits /GC–AG/ termini, while a tiny fraction exhibits /AC–AT/ termini (U12-type introns). U2-type and U12-type introns are spliced by distinct spliceosomes [2]. In addition to the precise recognition of exon–intron junctions among many possible pseudo-splice sites (that is, intronic nucleotides matching the splice site consensus) and the splicing of introns, the spliceosome also has to integrate this with other steps in RNA processing, such as capping, cleavage, and polyadenylation [6]. A picture emerges in which the control of gene expression is in part thought of as a network of interactions between transcription and RNA processing, export, and transcript quality control [1].

One gene, different messages. The splicing pattern of many pre-mRNAs is variable: different splice sites may be used as alternatives, giving rise to multiple alternatively spliced (AS) mRNA isoforms, and thus producing mature mRNAs and ultimately polypeptides that can be highly similar or markedly different while originating from the same locus [2–4]. Detailed molecular studies of regular and disease-

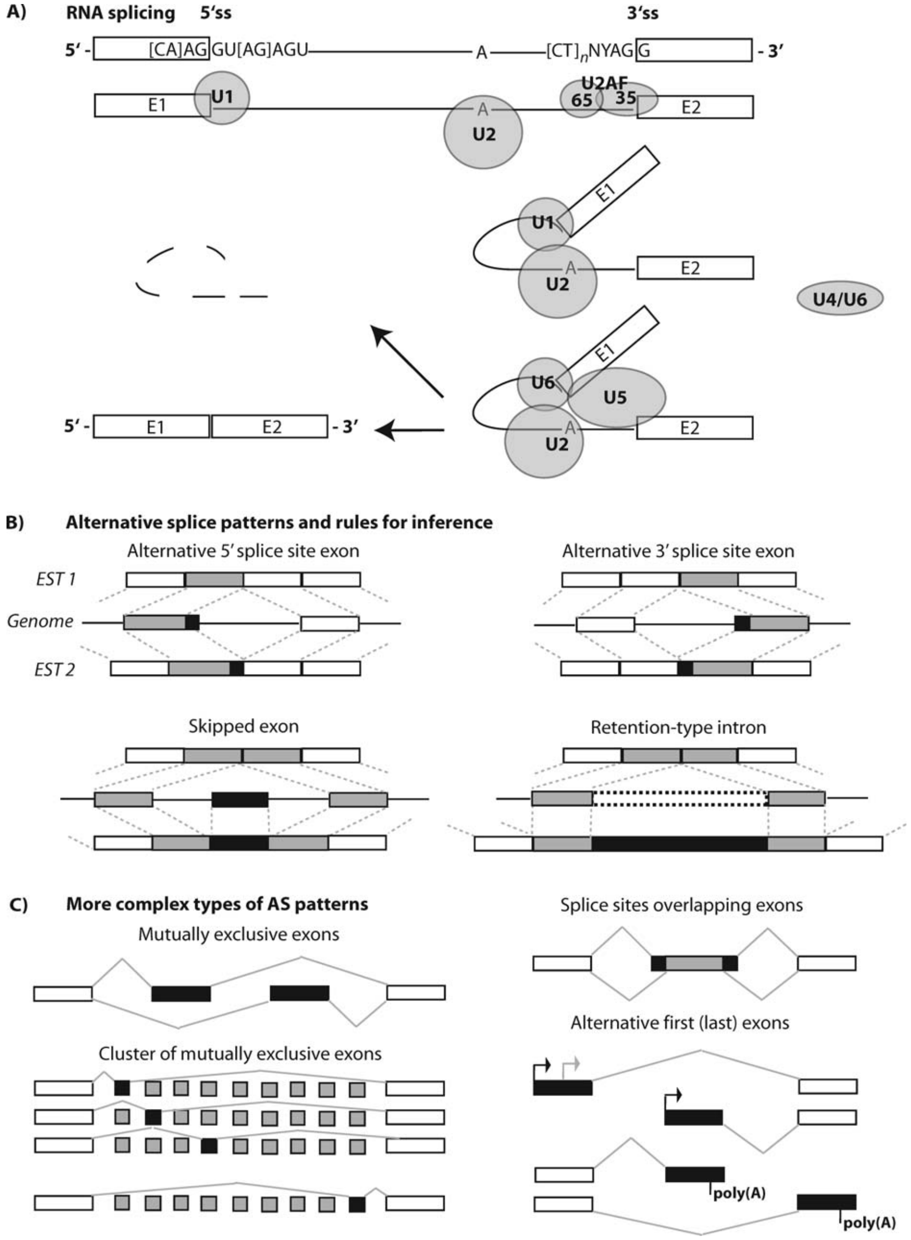
Editor: Fran Lewitter, Whitehead Institute, United States of America

Citation: Holste D, Ohler U (2008) Strategies for identifying RNA splicing regulatory motifs and predicting alternative splicing events. PLoS Comput Biol 4(1): e21. doi:10.1371/journal.pcbi.0040021

Abbreviations: 3′ss, 3′ splice site; 5′ss, 5′ splice site; A5E, alternative 5′ss exon; A3E, alternative 3′ss exon; AS, alternative splicing or alternatively spliced; cDNA, complementary DNA; ESE (ESS), exonic splicing enhancer; EST, expressed sequence tag; NMD, nonsense-mediated mRNA decay; SE, skipped exon; SJ, splice junction.

Dirk Holste is with the Research Institute of Molecular Pathology and the Institute of Molecular Biotechnology of the Austrian Academy of Sciences, Vienna, Austria. Uwe Ohler is with the Institute for Genome Sciences & Policy, Department of Biostatistics & Bioinformatics, Department of Computer Science, Duke University, Durham, North Carolina, United States of America.

* To whom correspondence should be addressed. E-mail: uwe.ohler@duke.edu (UO), holste@imp.ac.at (DH)


A)
RNA splicing
5'ss
3'ss
5' - [CA]AG GU[AG]AGU A [CT]nNYAG G - 3'
E1
U1
A
U2
U2AF
65
35
E2
U4/U6
U6
U5
5' - E1 E2 - 3'
B)
Alternative splice patterns and rules for inference
Alternative 5' splice site exon
Alternative 3' splice site exon
EST 1
Genome
EST 2
Skipped exon
Retention-type intron
C)
More complex types of AS patterns
Mutually exclusive exons
Splice sites overlapping exons
Cluster of mutually exclusive exons
Alternative first (last) exons
poly(A)
poly(A)

doi:10.1371/journal.pcbi.0040021.g001

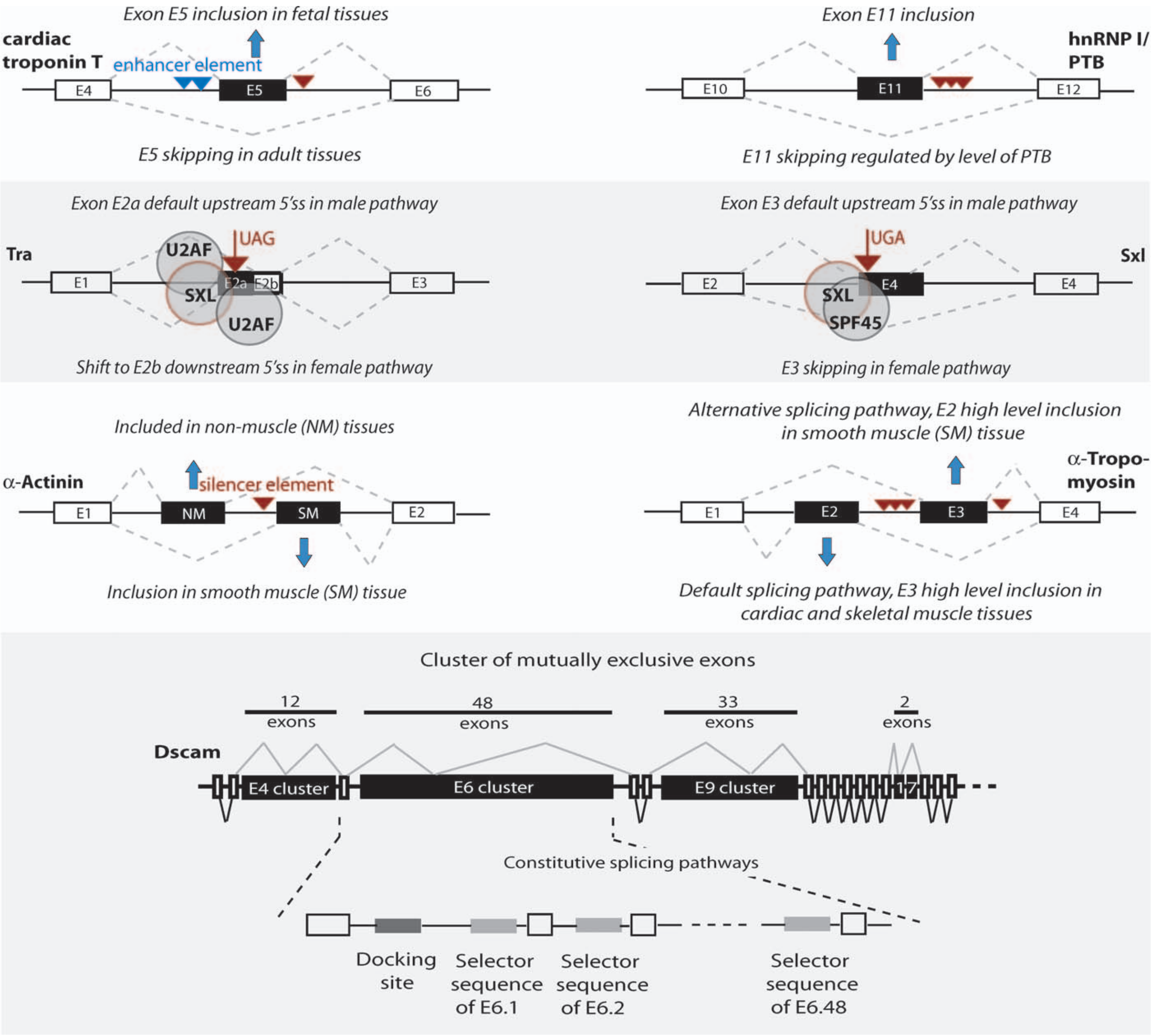
Figure 1. Basic Steps of Pre-mRNA Splicing and Patterns of Alternative Exons

(A) Five small nuclear ribonucleoproteins (snRNPs U1, U2, U4, U5, and U6), an auxiliary splicing factor (U2AF), and many other factors (not represented) organized in the human spliceosome execute the excision of introns. After the recognition of 5′ss, 3′ss, and branch point, respectively, by U1, U2AF, and U2, the intron is first cleaved at the 5′ss and subsequently at the 3′ss, mediated by U4/U6 and U5 (U1, U2, and U4 are detached later during the cycle). The intron remains in the nucleus and is degraded, while ligated exons are transported outside to the cytoplasm.

(B) AS events can be inferred by spliced alignments of mRNAs to genomic DNA (cf. Figure 3), indicated by dashed lines (AS part of exon colored in black), and commonly distinguished in terms of whether mRNA isoforms differ by skipping of an exon (SE), or whether isoforms differ in the usage of a 5′ss or 3′ss, producing an A5E or A3E, respectively. A fourth type, termed retention-type intron, occurs when two isoforms differ by the presence of an unspliced intron in one transcript that is spliced in the other.

(C) More complex types of AS forms can be constructed from canonical splice variants; different isoforms can also be the result of variations at the 5′- and 3′-terminus of transcripts, which are not necessarily due to AS.

associated genes have identified several hundred genes subject to AS. One powerful but not typical example of the possibilities opened up by the process of AS is given by the *D. melanogaster* gene *Dscam* (Figure 2), which has the potential to produce and express hundreds to thousands of alternative mRNA isoforms [7]. Computational analyses of available large datasets of spliced mRNAs to genomic DNA infer that a large number of mammalian genes (often estimated at more than 50%) produce messages that are consistent with variable splice site choices made during alternative pre-mRNA splicing [8,9]. In fact, the average number of detected isoforms per gene can reach such an extent that one can hardly distinguish the "alternative" transcript and might instead invoke the concept of a set of transcripts produced from a gene's locus [10]. Such whole-genome bioinformatics studies have added further lines of evidence to the



Exon E5 inclusion in fetal tissues
cardiac troponin T
enhancer element
E4
E5
E6
E5 skipping in adult tissues
Exon E11 inclusion
hnRNP I/PTB
E10
E11
E12
E11 skipping regulated by level of PTB
Exon E2a default upstream 5′ss in male pathway
Tra
UAG
U2AF
SXL
E1
E2a
E2b
E3
U2AF
Shift to E2b downstream 5′ss in female pathway
Exon E3 default upstream 5′ss in male pathway
UGA
Sxl
E2
SXL
E4
SPF45
E4
E3 skipping in female pathway
Included in non-muscle (NM) tissues
α-Actinin
silencer element
E1
NM
SM
E2
Inclusion in smooth muscle (SM) tissue
Alternative splicing pathway, E2 high level inclusion in smooth muscle (SM) tissue
α-Tropomyosin
E1
E2
E3
E4
Default splicing pathway, E3 high level inclusion in cardiac and skeletal muscle tissues
Cluster of mutually exclusive exons
12 exons
48 exons
33 exons
2 exons
Dscam
E4 cluster
E6 cluster
E9 cluster
17
Constitutive splicing pathways
Docking site
Selector sequence of E6.1
Selector sequence of E6.2
Selector sequence of E6.48

doi:10.1371/journal.pcbi.0040021.g002

Figure 2. Selection of Splice Patterns of Known Alternative Exons

Selection of splice patterns of known alternative exons of *Tra*, *Sxl*, and *Dscam* genes in *D. melanogaster* [3,4], and α-*Actinin*, α-*Tropomyosin*, *Troponin-T*, and *PTB* in *H. sapiens* [71]. Exon skipping is the predominant AS event in many metazoans and, e.g., has been shown to be involved in tissue- and developmental stage–specific regulation, as well as autoregulation (*PTB*) [72]. AS products of pre-mRNAs expressed from *Tra* and *Sxl* genes are involved in the pathway of somatic sex fate in *D. melanogaster*, which is regulated by altogether five AS genes at the top of the determination cascade [4]. The "master gene" *Sxl* is expressed in female flies, where it acts as a negative regulator of splicing. AS of the *Dscam* gene is known for its theoretically large number of possible different AS products (~38,000 against ~14,000 *D. melanogaster* protein-coding genes), which are derived from four clusters of skipped exons. The regulation of one cluster includes so-called selector-docking sites, which are inverse complementary overlapping sites located in the most 5′-end intron (docking) and upstream of each skipped exon (selector) of this cluster, respectively [73].

occurrence and scope of AS, such that it is now considered to be critically contributing to the diversification of proteins expressed in different cell types and developmental stages. As splicing is critical to the viability of the cell, it is clear that nonphysiological splicing decisions can have pathological effects, and consequently splicing and AS are gaining interest as possible explanations for human genetic disorders [11]. Figure 2 displays splice variants of known AS genes, originating from exon-skipping, alternative 3′ss exons, or mutually exclusive exon splicing.

In addition to protein diversification, AS might have another function in the realm of gene regulation, by linking splicing to a downstream control mechanism, termed nonsense-mediated mRNA decay (NMD) [12]. Under this scheme, aberrant or deliberately produced mRNA isoforms that harbor shifts of the original reading-frame and hence lead with high probability to premature termination codons downstream of the frame shift, are candidate substrates for NMD and shutting down of protein synthesis. While computational studies have inferred a significant and large number of possible NMD target isoforms [13], first interrogations of splicing-sensitive micro-arrays and NMD mutants have so far failed to detect large support for a widespread utilization of this mechanism [14].

Computational challenges. As experimental systems and models for the regulation of AS have been steadily validated and refined, so have bioinformatics tools and computational models [15,16]. With the availability of complete genome sequences and comparative genomics, the identification of candidate sequence elements that can be evaluated for their activity in controlling gene expression has become a major challenge in computational molecular biology. In order to systematically address the level of complex control achieved by AS, experimental and computational large-scale studies have started to illuminate the extent, structure, and regulatory consequences of AS and its differential usages in mammalian genomes. Here, we focus on two selected aspects out of a large body of computational approaches, which have been at the center of several recent studies: starting with basic steps for data acquisition and splice patterns classification, we first present an overview and compare several methods for candidate splicing *cis*-regulatory element detection. Subsequently, we move toward the goal of predictive identification of AS events from genomic sequence. There are many additional aspects of AS worth mentioning, ranging from specific algorithms for spliced alignments to the population genetics of exon and intron evolution, and for these we would like to refer to additional recent excellent reviews [4,17–20].

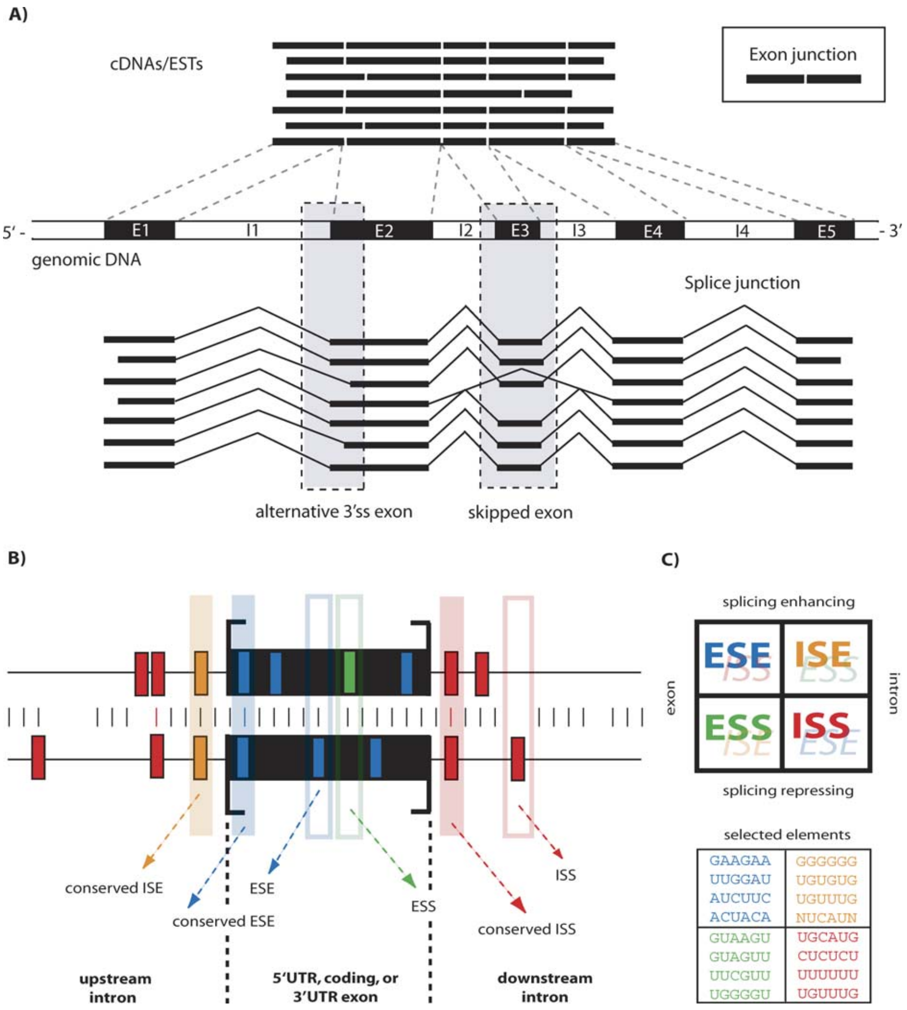
From Transcripts to Patterns of Alternative Splicing

The principal computational approach to identifying AS genes, and to infer individual alternative exon events or complete alternative isoform structures, relies on the comparison of available transcript data to assembled genomes and known gene loci. For this framework to work well, one needs available complete and annotated genomes, large collections of transcribed sequences acquired under various cellular contexts (e.g., different cell or tissue types), and reliable and efficient algorithms for sequence alignment. In an initial step, large-scale alignments of transcripts to genomic DNA are conducted using a variety of systems (a selection is listed in Table S1). The genomic sequence is usually of high-end quality, whereas transcribed sequences come in two different flavors: 1) complementary DNAs (cDNAs), which often produce a single or at least very limited number of possible genomic matches; and 2) expressed sequence tags (ESTs) or shorter reads from massively parallel sequencing, which can produce a considerably larger number of possible matches. ESTs are sequenced in a single pass and are therefore available in large numbers, but often quite error-prone especially toward the ends of a read.

At the heart of spliced-alignment algorithms are often dynamic-programming approaches: given a transcribed sequence (the mRNA), provide an alignment of a second contiguous sequence (the genomic DNA) to it that is allowed to be interrupted by long gaps which correspond to spliced-out introns (Figure 3A). Standard gap opening/extension penalties are not appropriate in this context; rather, gap penalties should be based on intron length distributions, and gaps should preferentially appear at positions that correspond to splice sites. Practically, such an alignment is most feasible when the contiguous sequence is restricted to genic DNA (e.g., ENSEMBL- or similarly known annotated genes). In the context of millions of available EST sequences, some systems also use shortcuts to avoid the often-prohibitive quadratic runtime complexity. The work flow directing the GENOA system [10] can serve as a practical example and is similar to those in many databases: i) find candidate matches of identity between (repeat-masked filtered) cDNA sequences and genomic DNA; ii) determine spliced-alignments of significant matches of cDNAs to gene loci; iii) find matches of EST to successfully aligned cDNAs; and iv) splice-align significant matches of ESTs to previously cDNA-aligned gene loci. Afterward, quality filters are typically applied, e.g., on the number of hits of cDNAs/ESTs, percent cDNA/EST sequence aligned and sequence identity, minimum exon and intron sizes, maximum intron size, canonical splice sites, or the exclusion of genes that are subject to frequent DNA rearrangements (e.g., immunoglobulin genes).

After the genome-wide alignment, the annotation of constitutive and alternative exons is the next step. All transcripts aligned to a gene's locus are scanned for AS events, to identify alternative isoforms and which exons or introns they affect. In this framework, constitutive exons are the "default" status of exons, and this status remains unless specified conditions for annotation as an alternative exon are met. To this end, one can order observed splice junctions (SJs) and their frequency of occurrence, and construct an SJ matrix—with exons considered as "nodes" and SJs as "edges" connecting nodes. A traversal through this graph can capture different AS patterns [21]. This scheme is exon-centric, in that splice patterns are individually evaluated for each exon. Individual events are commonly categorized in four canonical patterns (Figure 1B), referred to as "skipped exon" (SE, or cassette exon), alternative 5′ss (A5Es) or alternative 3′ss exons (A3Es), or retention-type intron. These descriptions are not necessarily exclusive, and an exon can make several alternative splice site choices. Exon-centric schemes can only detect canonical events, but can be extended to capture more complex events, such as mutually exclusive exons or clusters of skipped exons (Figure 1C). In order to pair-wise compare isoforms generated from one gene against one another, a possible heuristic solution is to capture the total number of SJs that differ between two transcripts and normalize it to the total number of SJs, within a region where both transcripts overlap genomic DNA [22]. Storing ingoing and outgoing edges in the SJ matrix allows for the construction of complete isoforms as a representative path through the graph [16].

Computational analyses indicate that AS predominantly generates SE events in both human and mouse transcriptomes [22], and likely generally across vertebrates [23]. The frequency of SE events is followed by A5E and A3E events, which in turn are followed by retention-type introns, overlapping exons (simultaneous occurrence of A5E and A3E events), and mutually exclusive exons. When accumulating splice variants and monitoring the respective frequency of alternatives, one often obtains a bimodal distribution of the percentage of events, in particular for the inclusion and exclusion of SEs, and the two variants are in such cases referred to as "major" and "minor" isoforms depending on their frequency. Expectedly, the availability of a larger



doi:10.1371/journal.pcbi.0040021.g003

Figure 3. From Sequences to Patterns and Functional Elements

(A) AS events can be computationally inferred by spliced-sequence alignments of complete or partial mRNAs to genomic DNA. A selection of available algorithms and software is listed in Table S1. The sketch shows seven mRNAs with indicated exon junctions (for visual guidance only), the primary transcript structures of which are to be inferred from alignments to genomic DNA (the order of the mRNAs above and below the genomic DNA is the same). In the example shown, the set of mRNAs aligns to five exons (E1 to E5), and the data are consistent with two AS events: E2 alternative 3′ss splicing, and E3 skipping (skipped in the fourth mRNA from the top).

(B) Splicing-regulatory elements are distinguished depending on their location (exon or intron) and their mode of action (enhancing or silencing): 1) exonic splicing enhancer (ESE) elements; 2) exonic splicing silencer (ESS) elements; 3) intronic splicing enhancer (ISE) elements; and 4) intronic splicing silencer (ISS) elements. One can subclassify these elements whether they carry protein-coding information, act in the context of 5′ss and/or 3′ss, or are sequence-conserved across species (indicated by the presence of vertical colored bars).

(C) Often, ESE, ESS, ISE, and ISS elements do not act independently of their sequence context, but can assume antagonistic functions (enhancing versus silencing) in splicing. The color-coded example sequence elements are taken from the literature [27,28,32,38].

number of cDNAs and ESTs from a gene increases the chance of observing alternative isoforms of that gene, so the proportion of AS genes will tend to increase with increasing transcript coverage of genes. Probabilistic and sampling strategies have been discussed to circumvent or correct for this bias [21,22,24].

Signals for Splicing Specificity and Strategies for Their Identification

A comparison of general splicing signals (5′ss, 3′ss, and branch point) between baker's yeast, *C. elegans*, *A. thaliana*, and humans shows that the information content of these signals is less and less preserved with an increasing number of introns [25,26], and in higher eukaryotes is often insufficient to ensure correct RNA splicing. The growing degeneracy, in turn, opens up the possibility for making alternative splice sites choices, and additional signals become necessary [27], in particular when weak splice sites are involved [28]. It is known that the choice of a splice site can be affected by a number of features, including exon and flanking intron size, splice site strength, splicing regulatory elements, interspersed repeat content, mRNA secondary structure, or RNA editing. Splicing-specific elements function in context as exonic/intronic splicing enhancers (ESE/ISE elements) or silencers (ESS/ISS elements), and alter splice site choices by recruiting positive or negative *trans*-acting regulatory factors (Figure 3B). All elements are more or less ubiquitous in constitutive and alternative exons, short (~6–10 nucleotides long, with some longer exceptions), and present in the majority of exons or introns [29,30].

While there is presently no validated large set of such elements, tested in various standard splicing contexts, ESE and ESS elements are currently the best characterized [17,31]. ESE elements often recruit arginine/serine dipeptide-rich (SR) proteins, which themselves recruit other spliceosomal components via protein–protein interactions. The family of nuclear heterogeneous RNPs (hnRNPs) characterizes another class of splicing factors that often antagonizes members of the SR protein family and are thought to be recruited by ESS elements. Biochemical investigations have further revealed highly specific factors that bind to splicing regulatory elements, including members of the CELF [32], NOVA [33,34], PTB [35], and FOX [36] families of proteins. FOX-1, for example, is an RNA-binding protein expressed in brain, heart, and skeletal muscle tissues, and binds to the UGCAUG motif [37]. Interestingly, this motif has been computationally identified downstream of the 5′ss region, by searching AS genes specifically expressed in the human brain displaying exon-skipping events [38].

Early indications for the role of splicing regulatory elements were obtained, e.g., from disease-associated studies where a chance disruption pointed to the presence of functional sites, and often to evolutionary conservation around sites. Subsequently, several systematic computational and/or experimental assays have been developed to identify ESE, ESS, and other elements, and they can be grouped into the following main classes:

1. A **functional SELEX** (systematic evolution of ligands by exponential enrichment) approach was employed to search for ESE elements [39]. Using repeated rounds of selection, a library of random oligonucleotides was inserted into a minigene construct of known splicing behavior, replacing an authentic ESE. An entire pool of constructs, each with a different random oligonucleotide, was transcribed and spliced to produce a pool of mRNAs. For the next round, products of spliced mRNAs were amplified and used to construct a new generation of minigenes, thus starting another cycle. After several rounds, the "fittest" sequences with the desired splicing activity were sufficiently enriched and could then be extracted by sequencing. Subsequently, motif searches and/or multiple alignments are used to construct scoring matrices, e.g., for ASF/SF2, SC35, SRp40, or SRp55 [39,40].

2. A **splicing reporter system** was developed in [27] to systematically screen for ESS elements. To this end, a three-exon (E1-E3) minigene construct was designed, where first and last exons together encode the complete green-fluorescent protein (GFP) and the test exon (E2) is located between E1 and E3. Two states were of interest: 1) when E2 was included into the mature transcript, the resulting protein was not functional; 2) when E2 was skipped, E1 and E2 formed functional GFP. Using a library of random oligonucleotides, which were individually inserted into E2, test constructs were transfected into cultured cells. The cells were then automatically screened for signals of GFP-expression by fluorescent-activated cell sorting (FACS), and GFP-active cells could then be extracted for sequencing.

3. **Computational identifications of candidate splicing-regulatory elements** in exons and introns on a genome-wide scale have been conducted, and they can be grouped into searches for elements involving one species, or comparative searches in genomes of related species (Table S2). RESCUE-ESE elements [28] were identified by statistical analyses of exons, flanking intron regions, and splice site composition. Building on the observation that ESEs can compensate for weaker 3′ss and/or 5′ss of constitutive exons, ~240 human RESCUE-ESE motifs were predicted in a large exon set, by selecting hexamers that were enriched in exons against introns and weak against strong splice site scores. To further validate RESCUE-ESE–predicted motifs, a population genetics strategy (VERIFY) was developed [41] to assess the extent of purifying selection on functional sequences. Using a large collection of human single-nucleotide polymorphisms (SNPs), VERIFY estimated that about one-fifth of mutations disrupting RESCUE-ESE elements were eliminated by selection.

PESE/PESS elements (putative exonic splicing enhancers/silencers) were similarly identified, but come with a different flavor by avoiding any potential bias resulting from codon usage [42]. Here, the frequency of occurrences of oligonucleotides in noncoding exons was contrasted against pseudo-exons and 5′-UTRs of intronless genes. Oligonucleotides that were sufficiently overrepresented in noncoding exons were selected as PESE elements, while underrepresented ones were selected as PESS elements.

ESR (exonic splicing-regulatory) and ISR (intronic splicing-regulatory) elements were identified in comparative analyses [43,44]. To this end, the former approach used a comparison of the frequency of expected against observed codon pairs (hexamers), which were additionally highly conserved in the codon wobble positions between exons of orthologous *H. sapiens* and *M. musculus* genes. The latter approach focused on four-way conserved oligonucleotides in 400 nucleotides long intronic regions upstream and downstream of all flanking exons, by including the additional mammalian genomes of *C. familiaris* and *R. norvegicus*. Statistically enriched oligonucleotides were retained and clustered into groups based on their conservation. A related approach to search for regulatory elements in two nematodes, *C. elegans* and *C. briggsae*, was used in [45].

Out of a total of 4,096 different hexamers, the above searches predicted elements that span a range between several hundred to more than 50% of all hexamers, some of which overlap to a large extent. This raises the question of how much information is already contained in these sets, and which splicing-regulatory elements possibly remain to be predicted. An approach to group-predicted motifs was developed in [46], based on compositional differences ("distances") between elements. It inferred both ESE and ESS (NI-ESE/NI-ESS) elements from a compendium of RESCUE-ESE, FAS-ESS, and PESE/PESS elements, by using the sequence similarity to known ESE and ESS hexamers and a discriminating function to group between positive, negative, and splicing-neutral activity. Table S3 shows an all-against-all sequence comparison between the different sets of predicted regulatory elements, and Dataset S1 collects all these splicing *cis*-regulatory elements discussed above.

Several lines of evidence suggest that the influence of *cis*-regulatory elements exerted on splice site choices is context-dependent, and consequently the label "ESE" (or any other) is correct only in so far as the context is considered. An ESE element may well act as negative regulator of splicing in another context, e.g., when inserted into flanking intron regions or other exons (cf. Figure 3C).

Computational Models and Methods for the Prediction of Alternative Splice Site Choices from Sequence

Phylogenetic conservation of alternative splicing. The identification of AS events has traditionally been based on the analysis of the diversity observed in transcribed sequence data. As gene expression is largely condition-specific, it is hard to predict when we will have generated and sequenced EST libraries under sufficiently different conditions, and sufficiently deep to arrive at a complete compendium of transcript diversity. This brings up an additional caveat: noise occurs both on the level of experiment as well as in the cell. Filters imposed on EST-based AS inference help to reduce experimental noise and eliminate unwanted contamination. Yet, given the considerable degeneracy of sequence signals for splicing, the machinery itself is inherently noisy and prone to (reproducible) errors, and cellular mechanisms such as NMD have evolved to eliminate erroneously spliced products. As it has been provocatively put, one may be able to observe the alternative use of any possible splice site if one just sequences enough ESTs [47]. The question therefore is how many AS events actually correspond to a specific function. In any case, widespread AS may serve as "evolutionary tunneling" [9,48] and provide an organism with a mechanism to quickly "explore" new isoforms, only few of which eventually become functional and get fixed. In a larger context, this connects to the question of the evolution of AS and the structure of eukaryotic genes per se [19,49].

To reduce spurious nonfunctional alternatives, conservation was initially used as a filter. More recently the focus of comparative genomics has shifted toward identifying AS events that are 1) conserved with respect to orthologous genes ("alternative conserved exons", or ACEs); 2) only alternatively spliced in one species; or 3) newly created alternative exons, which are absent in orthologous genes of other lineages (Figure 4A). Given the differences in primary datasets and protocols for the inference of AS, it is difficult to estimate the fraction of splicing-conserved exons of orthologous genes. Based on transcript-inferred and predicted ACEs, the proportion of ACEs shared between *H. sapiens/M. musculus* is estimated to be ~11% of all EST-derived skipped human exons [50], while other estimates determine about 50% or even higher levels of conservation [51,52]. Possible reasons for such differences point to a lack of standards across different datasets and databases, and could possibly be attributed to differences in cDNAs/ESTs, inference of constitutive and/or alternative exons, mixing of splice variants (in terms of alternative splice site usage, or high- and low-frequency inclusion), orthologous gene relationships, or stringency of comparative analyses.

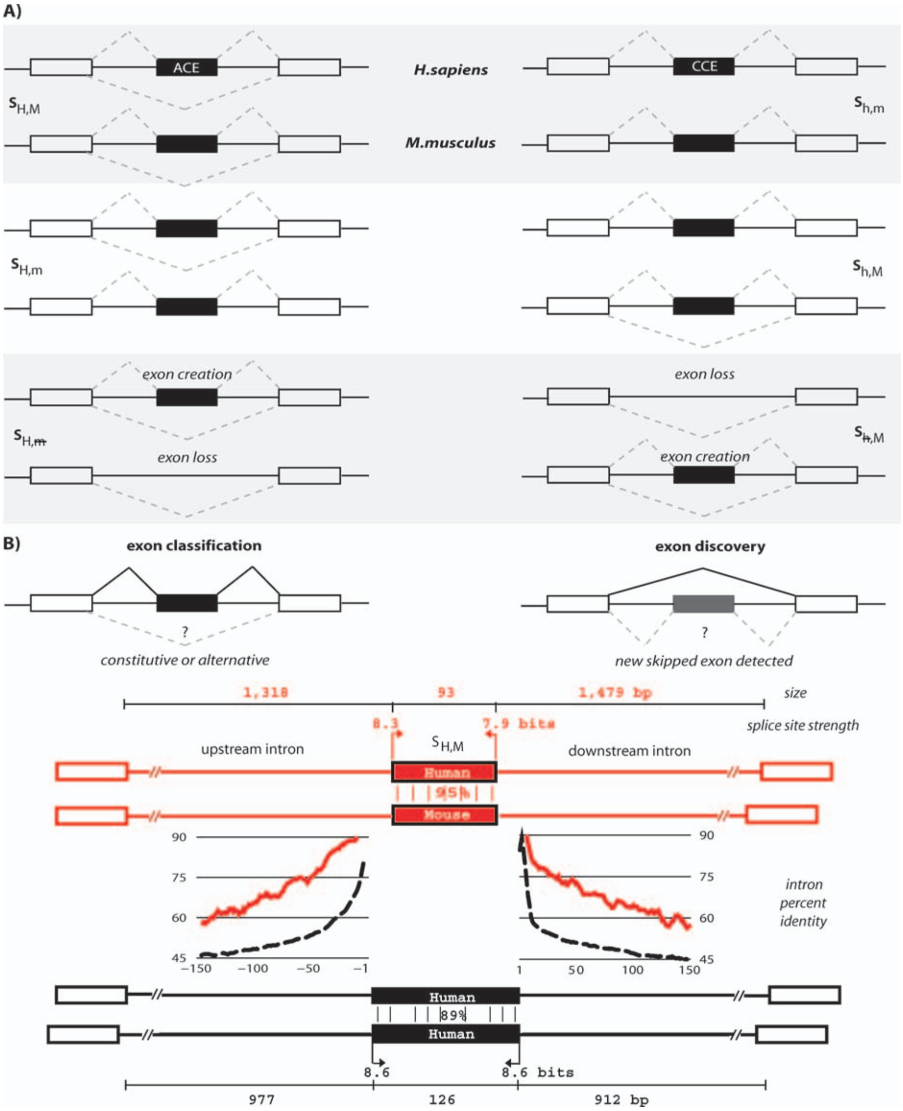
Ab initio prediction of alternative splicing events. In response to incompleteness and noise issues that come with transcript-derived isoforms, recent years have seen a number of approaches that aim at the direct "ab initio" identification of AS isoforms, by methods that solely rely on (comparative) sequence information of genomic DNA alone, without additional data such as expressed sequences or protein information. This is possible because exons affected by AS have different characteristics when compared to constitutive ones. In mammals, exon-skipping is the predominant type of AS, and sets of ACEs of orthologous human–mouse genes could be identified and analyzed for such functional characteristics [53]. Compared to constitutive conserved exons, ACEs are on average shorter, have weaker splice sites, and exhibit higher sequence-conservation of the exon body and flanking intron regions [53], lower ESE frequencies [40] (Holste, unpublished data), fewer SNPs, and are under higher natural selection pressure [47]. They are also more likely to preserve the reading frame and less likely to disrupt protein domains, are enriched in genes expressed in the brain, and in genes involved in transcriptional regulation, RNA processing, and development [50]. Several "non-transcript-based" algorithms have utilized these observations to identify new AS isoforms, and they can be grouped into two main classes addressing related but different problems:

1. **Exon *classification* algorithms** take known exons as input and classify them into "constitutive" or "alternative". Typically, a classifier is built on known examples of constitutive and AS exons, and uses sequence features extracted from the exon and surrounding introns. The approaches frequently work on known pairs of orthologous exons, as conservation features strongly contribute to the performance of the classifier. An exception is the single-species approach of [54], which was applied to the genome of *C. elegans*, and where the missing information from conservation is arguably offset by the simpler genome organization compared to mammals.

One of the first such algorithms used only exon and flanking intron conservation as features to predict that a given conserved exon was subject to any form of AS in two species of *Drosophila* [55]; despite the simplicity of the features, its specificity was more than 40%. At the same time, the first system to detect ACEs in the mammalian genomes combined thresholds of exon and intron conservation with exon size and frame conservation, and achieved a sensitivity of ~32% with virtually no false positives [56]. The sensitivity of these simple approaches can be dramatically improved by using a larger set of features, e.g., the presence of oligonucleotides corresponding to known or predicted sequence elements, combined with solid machine learning approaches that include the selection of significant features from the large set of motifs or the utilization of sparseness priors [50,54,57,58]. The ACESCAN system, for instance, reported an equal recognition rate (balanced sensitivity and specificity) of ~90%, while the specificity of newly predicted ACEs was ~70% in experimental validations [50].

2. **Exon *discovery* algorithms** take introns as input and parse them for the presence of hitherto unknown (and thus presumably mostly skipped) exons. This class has received somewhat less attention; the methods are generally based on gene finding–related approaches such as pair hidden Markov models (e.g., in the UNCOVER system [59]), which parse a pair of orthologous input sequences into segments with different functions and/or patterns of conservation, such as splice sites or coding triplets [60]. This approach has been extended to utilize multiple alignments of several species of *Drosophila* [61].

The exons in classification algorithms are known already, and this usually implies that the major isoform is exon inclusion; in comparison, newly discovered exons tend to be excluded in the majority of transcripts, which explains why they have not yet been annotated. Figure 4B exemplifies the



doi:10.1371/journal.pcbi.0040021.g004

Figure 4. Conservation of AS across Species

(A) Splice patterns of exons of pairs of orthologous genes can be classified into four "pattern conservation" categories, demonstrated here for SE events in *H. sapiens/M. musculus*: both exons are constitutively spliced ($S_{h,m}$)—known as "constitutive conserved exons" (CCEs); the exon of the human gene is alternatively spliced, the mouse one constitutively ($S_{H,m}$); the exon of the mouse gene is alternatively spliced, the human one constitutively ($S_{h,M}$); both exons are alternatively spliced ($S_{H,M}$)—known as "alternative conserved exons" (ACEs). In addition, one can define two "gain/loss" categories as: the exon of the human gene is alternatively spliced, the mouse exon is absent ($S_{H,\cancel{m}}$); the exon of the mouse gene is alternatively spliced, the human exon is absent ($S_{\cancel{h},M}$).

(B) AS events can be successfully predicted ab initio, that is, from genomic sequence alone. The figure refers to SEs in particular, but results for other classes of AS events have also been described. Published approaches address one or both of two problems: exon classification and exon discovery. Constitutive and alternative exons show different characteristics, which can be used as features for nontranscript methods. Example features include length of the AS exon and surrounding intron; strength of the splice sites; the level of conservation in the exon and surrounding introns; and coding-typic conservation patterns in exon sequences. In addition, some approaches utilize the occurrence of specific sequence features corresponding to splicing regulatory elements (parts of the figure adapted from [50]).

differences of the two approaches, as well as some of the typical features used by the classifiers.

Current algorithms mostly deal with the case of SEs; however, other AS types have been tackled as well. In addition to identifying new ACEs, the UNCOVER model also allows us to predict cases of conserved coding *H. sapiens/M. musculus* retention-type introns [59]. An early approach focused on the identification of new partners in a mutually exclusive pair of SEs [62]; these exons often arise from local duplications and therefore share considerable sequence similarity. Additional models use protein domain information to identify which newly predicted AS isoforms, generated by exon skipping and/or retention-type introns, would generate known protein domains [63,64].

How much do these exons contribute to the overall variability in gene structures? ACESCAN (an exon classification method) predicted ~4,000 exons to be ACEs in the human genome [50]; UNCOVER (an exon discovery algorithm) predicted ~50 new splicing-conserved SEs in human ENCODE regions [59] and ~8,500 exons genome-wide, which are annotated neither in *H. sapiens* nor in *M. musculus*. When coupled with ACESCAN, more than 6,000 of these passed as putative AS candidates (Ohler and Yeo, unpublished data). A similar survey used the phyloHMM EXONIPHY to arrive at ~700 new AS exons—an order of magnitude less, possibly due to initial conservation requirements in several additional mammalian species [65]. Concerning retention-type introns in coding regions, UNCOVER predicted the surprisingly low number of two-dozen conserved retention-type introns across *H. sapiens*–*M. musculus* orthologous protein-coding genes. In accord with this small number, Hiller et al. [63] identified 65 coding retention-type introns, but with the majority involving noncanonical splice sites not modeled in UNCOVER. A direct comparison of transcript-derived retention-type events was also indicative of low conservation (Nostrand, Holste, and Burge, unpublished data). Overall, Sorek et al. estimated that ~7% of coding exons undergo some type of AS conserved between *H. sapiens* and *M. musculus* [65]. Even if each of these variants would affect a different gene, this identifies a considerably sized gap between EST-observed and species-specific AS events on the one side, and conserved and presumably functional AS events on the other side.

None of these algorithms predicts complete mRNA isoforms; rather, they provide the building blocks of these by identifying the parts of a gene susceptible to different mechanisms of AS. Rare instances of ab initio gene-finding algorithms allow for predicting several complete alternative gene structures in the input sequence [66,67]; these algorithms are able to enumerate possible gene structures, but do not use information of functional splicing *cis*-elements or *trans*-factor concentrations to arrive at AS isoforms which are actually produced. A "splicing simulator" would take as input the sequence of a pre-mRNA and be able to automatically predict which isoforms exist and, with additional context information such as the expression levels of all splicing factors, how frequently they are generated. Compared to computational gene finders, which strongly rely on statistical properties of reading frame, coding content, and phylogenetic conservation [67], such an approach would only make use of the information the cell has available at the time of splicing in the nucleus. In practice, such splicing simulators are still in early stages of development, but preliminary successful results have already been achieved with approaches that combine models for splice sites explicitly [27] or implicitly [68] with other splicing regulatory motifs such as ESE and ESS elements.

Resources

Finally, we want to point to resources available to researchers who wish to obtain a deeper understanding of the transcript variability within genes of interest, and how this variability is generated and regulated on the level of RNA splicing.

Value-added databases. Databases for recording types of AS have been designed and operated for some time now, and they can be grouped into two approaches: 1) based on searches of published research [69,70]; and 2) automated large-scale comparisons of transcript sequences. Broadly speaking, the first approach emphasizes the manual curation and focuses on the "specificity" of (authentic) AS events, while computational approaches have their focus on "sensitivity" as well. Currently, the pipelines to annotate gene structures in these databases (and even more so in general-purpose genome browsers) are heavily driven by EST and homology evidence, and do not provide information on splicing regulatory elements or ab initio–predicted AS events or mRNA isoforms. AS databases that begin to include such information are becoming available and provide a more comprehensive picture of splicing and its regulatory elements (Table S4).

Bioinformatics software and Web servers. Several of the systems discussed here are accessible through a Web server or available as download for local analyses (cf. Table S5 for a selected overview). ■

Supporting Information

Table S1. Algorithms for Spliced-Sequence Alignments To Infer Primary Transcript Structures

Found at doi:10.1371/journal.pcbi.0040021.st001 (29 KB PDF).

Table S2. Overview of Comprehensive Motif Searches for Splicing-Regulatory Sequence Elements

Found at doi:10.1371/journal.pcbi.0040021.st002 (42 KB PDF).

Table S3. Comparison of the Different Classes of Predicted Regulatory Elements Described in Table S2

Found at doi:10.1371/journal.pcbi.0040021.st003 (45 KB PDF).

Table S4. Databases for Alternative Splicing

Found at doi:10.1371/journal.pcbi.0040021.st004 (59 KB PDF).

Table S5. Algorithms for Ab Initio Detection of Alternative Splicing Events

Found at doi:10.1371/journal.pcbi.0040021.st005 (32 KB PDF).

Dataset S1. Collection of All *cis*-Regulatory Elements with Predicted Splicing Enhancing or Silencing Activity Reviewed in This Article

Found at doi:10.1371/journal.pcbi.0040021.sd001 (1.4 MB XLS).

Acknowledgments

We thank G. Yeo and M. Stadler for helpful comments on the manuscript.

Author contributions. DH and UO wrote the paper.

Funding. This research was supported in part by the National Science Foundation (NSF), under grant PHY05-51164, during a stay at the Kavli Institute for Theoretical Physics, University of California Santa Barbara (DH, UO).

Competing interests. The authors have declared that no competing interests exist.

References

1. Maniatis T, Reed R (2002) An extensive network of coupling among gene expression machines. Nature 416: 499–506.
2. Burge C, Tuschl T, Sharp P (1999) Splicing of precursors to mRNAs by the spliceosomes. In: Gesteland R, Cech T, Atkins J, editors. The RNA world. 2nd edition. Cold Spring Harbor (New York): Cold Spring Harbor Laboratory Press. pp. 525–560.
3. Lopez AJ (1998) Alternative splicing of pre-mRNA: Developmental consequences and mechanisms of regulation. Annu Rev Genet 32: 279–305.
4. Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 72: 291–336.
5. Jurica MS, Moore MJ (2003) Pre-mRNA splicing: Awash in a sea of proteins. Mol Cell 12: 5–14.
6. Nilsen TW (2003) The spliceosome: The most complex macromolecular machine in the cell? Bioessays 25: 1147–1149.
7. Graveley BR (2001) Alternative splicing: Increasing diversity in the proteomic world. Trends Genet 17: 100–107.
8. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
9. Modrek B, Lee CJ (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet 34: 177–180.

10. Holste D, Huo G, Tung V, Burge CB (2006) HOLLYWOOD: A comparative relational database of alternative splicing. Nucleic Acids Res 34: D56–D62.
11. Faustino NA, Cooper TA (2003) Pre-mRNA splicing and human disease. Genes Dev 17: 419–437.
12. Lejeune F, Maquat LE (2005) Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. Curr Opin Cell Biol 17: 309–315.
13. Lewis BP, Green RE, Brenner SE (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A 100: 189–192.
14. Pan Q, Saltzman AL, Kim YK, Misquitta C, Shai O, et al. (2006) Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev 20: 153–158.
15. Lareau LF, Green RE, Bhatnagar RS, Brenner SE (2004) The evolving roles of alternative splicing. Curr Opin Struct Biol 14: 273–282.
16. Lee C, Wang Q (2005) Bioinformatics analysis of alternative splicing. Brief Bioinform 6: 23–33.
17. Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: Exonic mutations that affect splicing. Nat Rev Genet 3: 285–298.
18. Matlin AJ, Clark F, Smith CW (2005) Understanding alternative splicing: Towards a cellular code. Nat Rev Mol Cell Biol 6: 386–398.
19. Ast G (2004) How did alternative splicing evolve? Nat Rev Genetics 5: 773–782.
20. Xing Y, Lee C (2006) Alternative splicing and RNA selection pressure—Evolutionary consequences for eukaryotic genomes. Nat Rev Genet 7: 499–509.
21. Kan Z, States D, Gish W (2002) Selecting for functional alternative splices in ESTs. Genome Res 12: 1837–1845.
22. Yeo G, Holste D, Kreiman G, Burge CB (2004) Variation in alternative splicing across human tissues. Genome Biol 5: R74.
23. Kim E, Magen A, Ast G (2007) Different levels of alternative splicing among eukaryotes. Nucleic Acids Res 35: 125–131.
24. Brett D, Pospisil H, Valcarcel J, Reich J, Bork P (2002) Alternative splicing and genome complexity. Nat Genet 30: 29–30.
25. Lim LP, Burge CB (2001) A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci U S A 98: 11193–11198.
26. Irimia M, Penny D, Roy SW (2007) Coevolution of genomic intron number and splice sites. Trends Genet 23: 321–325.
27. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, et al. (2004) Systematic identification and analysis of exonic splicing silencers. Cell 119: 831–845.
28. Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. Science 297: 1007–1013.
29. Zheng CL, Fu XD, Gribskov M (2005) Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. Rna 11: 1777–1787.
30. Wang J, Smith PJ, Krainer AR, Zhang MQ (2005) Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. Nucleic Acids Res 33: 5053–5062.
31. Zheng ZM (2004) Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. J Biomed Sci 11: 278–294.
32. Ladd AN, Cooper TA (2002) Finding signals that regulate alternative splicing in the post-genomic era. Genome Biol 3: reviews0008.
33. Jensen KB, Dredge BK, Stefani G, Zhong R, Buckanovich RJ, et al. (2000) Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. Neuron 25: 359–371.
34. Grabowski PJ (2000) Genetic evidence for a Nova regulator of alternative splicing in the brain. Neuron 25: 254–256.
35. Rahman L, Bliskovski V, Reinhold W, Zajac-Kaye M (2002) Alternative splicing of brain-specific PTB defines a tissue-specific isoform pattern that predicts distinct functional roles. Genomics 80: 245–249.
36. Jin Y, Suzuki H, Maegawa S, Endo H, Sugano S, et al. (2003) A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. EMBO J 22: 905–912.
37. Nakahata S, Kawamoto S (2005) Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities. Nucleic Acids Res 33: 2078–2089.
38. Brudno M, Gelfand MS, Spengler S, Zorn M, Dubchak I, et al. (2001) Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. Nucleic Acids Res 29: 2338–2348.
39. Liu HX, Zhang M, Krainer AR (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev 12: 1998–2012.
40. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR (2003) ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res 31: 3568–3571.
41. Fairbrother WG, Holste D, Burge CB, Sharp PA (2004) Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol 2: e268. doi:10.1371/journal.pbio.0020268
42. Zhang XH, Chasin LA (2004) Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev 18: 1241–1250.
43. Goren A, Ram O, Amit M, Keren H, Lev-Maor G, et al. (2006) Comparative analysis identifies exonic splicing regulatory sequences—The complex definition of enhancers and silencers. Mol Cell 22: 769–781.
44. Yeo GW, Nostrand EL, Liang TY (2007) Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. PLoS Genet 3: e85. doi:10.1371/journal.pgen.0030085
45. Kabat JL, Barberan-Soler S, McKenna P, Clawson H, Farrer T, et al. (2006) Intronic alternative splicing regulators identified by comparative genomics in nematodes. PLoS Comput Biol 2: e86. doi:10.1371/journal.pcbi.0020086
46. Stadler MB, Shomron N, Yeo GW, Schneider A, Xiao X, et al. (2006) Inference of splicing regulatory activities by sequence neighborhood analysis. PLoS Genet 2: e191. doi:10.1371/journal.pgen.0020191
47. Baek D, Green P (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. Proc Natl Acad Sci U S A 102: 12813–12818.
48. Wang W, Zheng H, Yang S, Yu H, Li J, et al. (2005) Origin and evolution of new exons in rodents. Genome Res 15: 1258–1264.
49. Hong X, Scofield DG, Lynch M (2006) Intron size, abundance, and distribution within untranslated regions of genes. Mol Biol Evol 23: 2392–2404.
50. Yeo GW, Van Nostrand E, Holste D, Poggio T, Burge CB (2005) Identification and analysis of alternative splicing events conserved in human and mouse. Proc Natl Acad Sci U S A 102: 2850–2855.
51. Nurtdinov RN, Artamonova II, Mironov AA, Gelfand MS (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. Hum Mol Genet 12: 1313–1320.
52. Thanaraj TA, Clark F, Muilu J (2003) Conservation of human alternative splice events in mouse. Nucleic Acids Res 31: 2544–2552.
53. Sorek R, Ast G (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res 13: 1631–1637.
54. Ratsch G, Sonnenburg S, Scholkopf B (2005) RASE: recognition of alternatively spliced exons in C. elegans. Bioinformatics 21: i369–377.
55. Philipps DL, Park JW, Graveley BR (2004) A computational and experimental approach toward a priori identification of alternatively spliced exons. Rna 10: 1838–1844.
56. Sorek R, Shemesh R, Cohen Y, Basechess O, Ast G, et al. (2004) A non-EST-based method for exon-skipping prediction. Genome Res 14: 1617–1623.
57. Dror G, Sorek R, Shamir R (2005) Accurate identification of alternatively spliced exons using support vector machine. Bioinformatics 21: 897–901.
58. Sharan R, Myers EW (2005) A motif-based framework for recognizing sequence families. Bioinformatics 21: i387–393.
59. Ohler U, Shomron N, Burge CB (2005) Recognition of unknown conserved alternatively spliced exons. PLoS Comput Biol 1: 113–122.
60. Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge (United Kingdom): Cambridge University Press.
61. Allen JE, Salzberg SL (2006) A phylogenetic generalized hidden Markov model for predicting alternatively spliced exons. Algorithms Mol Biol 1: 14.
62. Letunic I, Copley RR, Bork P (2002) Common exon duplication in animals and its role in alternative splicing. Hum Mol Genet 11: 1561–1567.
63. Hiller M, Huse K, Platzer M, Backofen R (2005) Non-EST based prediction of exon skipping and intron retention events using Pfam information. Nucleic Acids Res 33: 5611–5621.
64. Leparc GG, Mitra RD (2007) Non-EST-based prediction of novel alternatively spliced cassette exons with cell signaling function in Caenorhabditis elegans and human. Nucleic Acids Res 35: 3192–3202.
65. Sorek R, Dror G, Shamir R (2006) Assessing the number of ancestral alternatively spliced exons in the human genome. BMC Genomics 7: 273.
66. Cawley SL, Pachter L (2003) HMM sampling and applications to gene finding and alternative splicing. Bioinformatics 19: II36–II41.
67. Flicek P, Brent MR (2006) Using several pair-wise informant sequences for de novo prediction of alternatively spliced transcripts. Genome Biol 7: S8 1–9.
68. Ratsch G, Sonnenburg S, Srinivasan J, Witte H, Muller KR, et al. (2007) Improving the Caenorhabditis elegans genome annotation using machine learning. PLoS Comput Biol 3: e20. doi:10.1371/journal.pcbi.0030020
69. Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O, et al. (2000) An alternative-exon database and its statistical analysis. DNA Cell Biol 19: 739–756.
70. Shah PK, Jensen LJ, Boue S, Bork P (2005) Extraction of transcript diversity from scientific literature. PLoS Comput Biol 1: e10. doi:10.1371/journal.pcbi.0010010
71. Spellman R, Rideau A, Matlin A, Gooding C, Robinson F, et al. (2005) Regulation of alternative splicing by PTB and associated factors. Biochem Soc Trans 33: 457–460.
72. Wollerton MC, Gooding C, Wagner EJ, Garcia-Blanco MA, Smith CW (2004) Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. Mol Cell 13: 91–100.
73. Graveley BR (2005) Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. Cell 123: 65–73.