Bioinformatics of alternative splicing and its regulation

Liliana Florea

Submitted: 3rd August 2005; Received (in revised form): 6th December 2005

Abstract

The sequencing of the human genome and ensuing wave of data generation have brought new light upon the extent and importance of alternative splicing as an RNA regulatory mechanism. Alternative splicing could potentially explain the complexity of protein repertoire during evolution, and defects in the splicing mechanism are responsible for diseases as complex as cancer. Among the challenges that rise in light of these discoveries are cataloguing splice variation in the human and other eukaryotic genomes, and identifying and characterizing the splicing regulatory elements that control their expression. Bioinformatics efforts tackling these two questions are just at the beginning. This article is a survey of these methods.

Keywords: *alternative splicing; gene indices; genome-based clustering; splice graph; splicing enhancers; splicing silencers*

INTRODUCTION

Alternative splicing of pre-mRNA has captured the attention of the genomics community as an important regulatory mechanism for modulating the gene and protein content in the cell. The discovery that many genes can produce multiple mRNA and protein isoforms, through the regulated selection of different combinations of exons for inclusion into the mRNA, changed irreversibly the outlook of many important genomics problems such as the annotation of genes and their regulation, and inspired hypotheses reaching far into the biomedical and evolutionary biology areas.

The phenomenon of alternative splicing was discovered in concept in the late 1970s [1], and was then verified experimentally in the 1980s [2], but the real revolution in alternative splicing occurred about the time of the sequencing of the human genome. The availability of large EST and mRNA data sets stimulated large-scale analyses which estimated that as many as 60% of the human genes undergo alternative splicing to create multiple transcript isoforms [3–5]. High levels of splicing have been estimated in other mammals recently sequenced, such as mouse and rat [6, 7]. Thus, alternative splicing appears to be the rule rather than the exception in gene expression. With the recent slashing of the number of human genes down to 25 000 [8], scientists have turned to alternative splicing as a mechanism to potentially explain how a large variety of proteins can be achieved with a relatively small number of genes [9], and ultimately to explain the paradox between gene content and the complexity of organisms. On a more immediate timescale, errors in the splicing mechanism and its regulation have been investigated as triggers in a number of diseases, including cancers [10–12], and such studies may lead to new tools for diagnostics and treatment [13–15].

Alternative splicing is the process through which different combinations of exons within a gene are spliced from the RNA precursor to be included in the mature mRNA depending upon the tissue, developmental stage and disease versus normal conditions of the cell. The resulting proteins may exhibit different and sometimes antagonistic functional and structural properties [13, 16], and may inhabit the same cell with the resulting phenotype being the balance between their expression levels [17]. From a gene structure standpoint, alternative splicing at internal exons manifests itself in four types

Liliana D. Florea, Department of Computer Science, George Washington Univeristy, Academic Center-Rm 714, Washington DC 20052. Tel: (202) 994-1057; Fax (202) 994-4875; E-mail: florea@gwu.edu

Liliana Florea is an Assistant Professor of Computer Science at the George Washington University. She develops algorithms and software for aligning cDNA and genomic sequences and for annotating genes and alternative splicing.

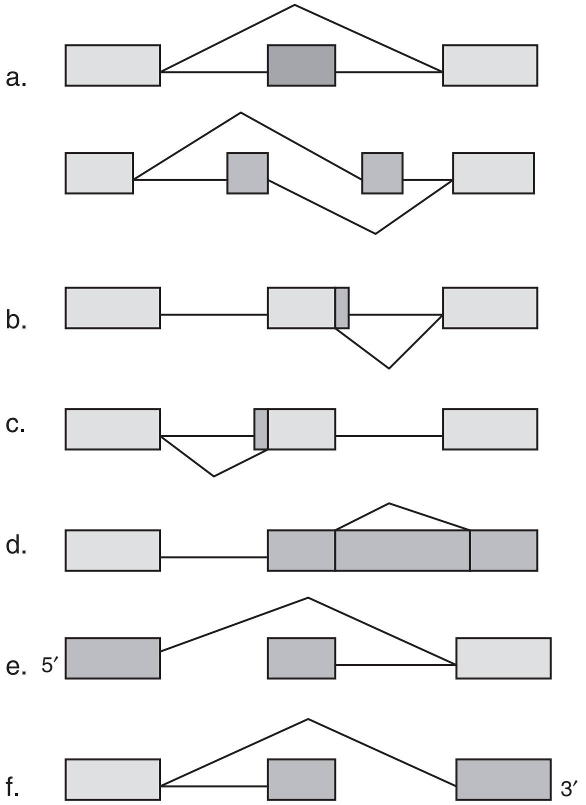

Figure 1: Types of alternative splicing events: (**a**). exon inclusion/exclusion; (**b**). alternative 3′ exon end; (**c**). alternative 5′ exon end; (**d**). intron retention; (**e**). and (**f**). 5′ and 3′ alternative untranslated regions (UTRs). The schematic representation shows exons as boxes, and introns as straight or segmented lines connecting the exons. Alternatively spliced elements (exons or portions of exons) are shown in dark grey, and those constitutively spliced in light color.

of events, depicted in Figure 1a–d: exon inclusion and/or exclusion, alternative 5′ exon end, alternative 3′ exon end, and intron retention. Alternative 5′ and 3′ untranslated regions (UTRs) are often encountered (Figure 1e–f), but in general are harder to identify and characterize due to noisy data and the intrinsic difficulty in identifying the gene boundaries. Consequently, they have not been reflected with the same prevalence in bioinformatics methods and literature.

The two most important reverberations of alternative splicing in bioinformatics and computational biology are in the areas of gene annotation and splicing regulation [18]. Traditionally, gene discovery was accomplished with a combination of *ab initio* and comparative methods designed to identify linear exon–intron models of genes along a genomic sequence [19]. Following the sequencing of the human genome, large-scale annotation projects at the major genomics centers (Ensembl [20], UCSC Genome Browser database [21], Celera Genome Browser and Otto annotation system [22]) typically employed a variety of prediction methods to produce 'evidence' for the gene, and used a '*combiner*' algorithm to consolidate the different lines of evidence into a representative gene model. Lastly, alternatively spliced forms were added to the automated annotation by human curators in their effort to improve the quality and completeness of the data set. Such sets were tedious to create and captured only partly the repertoire of splicing variation. More recently, the gene annotation landscape has changed to incorporate alternatively spliced transcripts or alternative splicing events as an integral part of the annotation process. There has recently been a concomitant increase in the number of repositories containing and cataloguing alternative splicing information.

Equally important is the question of what causes and/or controls the variation in splicing. Unlike transcriptional control exercised from the promoter, which modulates exclusively the amount of gene expression, the control of alternative splicing determines the abundance, structure and function of transcripts and encoded proteins from a gene, and can alter such salient properties of proteins as binding affinity, intracellular localization, enzymatic activity, stability and post-translational modifications [23]. Furthermore, exon selection in alternative splicing can be tissue, developmental stage or disease specific [13, 24], or can change in response to external stimuli such as receptor stimulation, cellular stress or changes in neuronal activity [25]. Thus, the regulation of alternative splicing presents a more versatile and finer granularity control than transcriptional regulation.

Most of the splicing regulation that is not part of the basal spliceosome function is known to be undertaken by families of splicing regulatory proteins. These splicing factors bind to signals in the vicinity of the exon and promote the exon's inclusion or exclusion by activating or inhibiting the function of the splice site. The number of classes and characteristics of these regulatory proteins and their RNA binding sites are relatively little known and are currently under active investigation.

In this article, we present an overview of bioinformatics methods to date to identify, characterize and catalogue alternative splicing and its regulation.

IDENTIFICATION OF SPLICE VARIANTS

While different tissues or cell types in an organism have roughly the same genome, their transcriptomes could be significantly different. The end goal of alternative splicing annotation is to identify and catalogue **all mRNA transcripts** in the cells of an organism at various stages, together with relevant information about their spatial and temporal expression and function. This is analogous to the gene annotation problem that has been in the cross hairs for the past two decades. Given the incomplete and fragmented nature of the data and the insufficient experimental characterization, this goal is currently difficult to achieve, if not impossible (specific challenges are listed in the forthcoming sections). A variation that is more readily attainable and can be just as powerful for practical purposes, such as designing diagnostic markers that can be tested *in vitro* via microarray and proteomic experiments [26] or *in silico* [27] is that of identifying and annotating **splice forms**, i.e. mutually exclusive splicing patterns inferred from partial cDNA and/or protein sequences, or exon-level **alternative splicing events**. Of particular interest are exons or combinations of exons that undergo different splicing patterns.

Bioinformatics

Approaches for identifying full-length splice variants or just splice forms typically involve the comparison of two or more expressed DNA or protein sequences of different isoforms to detect differences caused by insertions or deletions of genetic material. **Direct comparison** between the sequences of different cDNA and protein isoforms (Figure 2a) reveals the differences between the sequences compared, but does not characterize them in the context of the gene's structure. The observed difference can be an exon, part of an exon, or a set of consecutive exons and exon portions. In contrast, **comparisons of exon–intron structures** from cDNA- or protein-genomic spliced alignments clearly distinguish among the types of alternative splicing events (Figure 2b) and provide an extended genomic context in which the nearby splicing regulatory regions can be explored.

Sequence data

Computational approaches rely on cDNA and protein sequences collected in large repositories in GenBank and around the world. Database for expressed sequence tags (dbEST) [28] contains single-pass EST sequences from direct submissions to GenBank. RefSeq [29] is an NCBI effort to collect, review and curate full-length mRNA sequences from submissions or gene prediction projects. More recently, the Mammalian Gene Collection (MGC) project [30] was started as an NIH initiative to clone and sequence full-length open reading frames (ORFs) for human, mouse and rat genes. Protein sequences can be obtained from databases such as SwissProt, TrEMBL and PIR, currently united into the Universal Protein Knowledgebase UniProt [31]. The quality and characteristics of the data may differ significantly, and thus in judging the potential for identified differences to represent true alternative splicing events considerations about the type, quality and reliability of data are important.

Alignment tools

A number of specialized programs have been developed to align cDNA sequences to genomic sequences allowing for sequencing errors, polymorphisms and introns, such as EST_GENOME [32], Sim4 [33], Spidey [34], and GeneSeqer [35]. These programs were designed specifically to compare a cDNA with the restricted genomic range encompassing the gene. With the availability of whole genome sequences, new generation tools were developed to efficiently map large cDNA data sets to large chromosomal sequences and whole genomes. Examples include Blat [36], ESTmapper [37], MGAlignIt [38], and GMAP [39]. Given a cDNA sequence, each of these programs generates a spliced alignment of the cDNA and the genomic sequence. The alignment clearly marks the locations of exons and introns in the two sequences and gives additional information about the match, such as the predicted strand and alignment quality statistics. Although all of these programs have reached a high level of accuracy, challenges remain in dealing with non-canonical splice junctions, high EST sequencing error rates or specific types of sequencing errors, detecting small exons and/or large introns, and correctly determining the true location of the cDNA on the genome from among multiple paralogous matches.

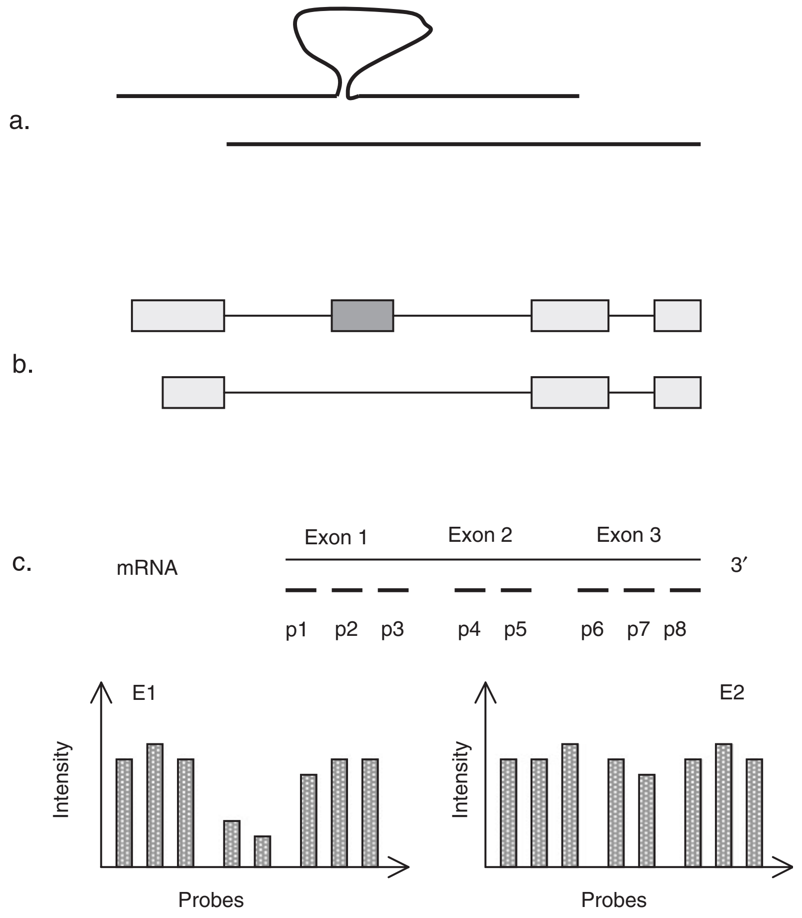

Figure 2: Bioinformatics methods for identifying alternative splicing. (**a**) Identification of alternative splicing events by direct comparison of cDNA sequences (**b**) comparison of exon–intron structures on the genome, (**c**) microarray experiments. In comparisons between pairs of cDNA sequences (**a**), alternatively spliced features appear as insertions in one sequence compared to the other. Comparison of exon–intron structures from spliced alignments of cDNAs on the genome (**b**) reveal the exons and introns undergoing alternative splicing. In (**c**), microarray GeneChip expression data using multiple probes per gene show potential alternative splicing events as groups of probes that are differentially expressed between the two experiments (probes p4 and p5 in exon 2). (**d**). Clustering and assembly of EST sequences into transcript-oriented groups and contigs. (**e**) EST sequences from the 5′ and 3′ ends of the gene are compared to identify compatible overlaps and then assembled into consensus sequences (TC I, TC 2) representing putative splice variants. Spliced alignments of cDNAs on the genome (EI–E5) are clustered along the genomic axis and consolidated into splice graphs. Vertices in the splice graph represent exons (a–h), arcs are introns connecting the exons consistently with the cDNA evidence, and a branching in the graph signals an alternative splicing event. Splice variants (VI–V4) are read from the graph as paths from a source vertex (with no 'in' arc) to a sink vertex (with no 'out' arc). In some systems, they are assigned scores based on the strength of the supporting evidence, based on which high-scoring candidates are selected for inclusion into the annotation and/or for validation experiments.

BIOINFORMATICS METHODS

Since identifying partial splice forms and alternative splicing events via differences between different isoforms is intuitive, subsequently we focus on the more complex task of annotating full-length alternatively spliced transcripts. Such methods include gene indices, which assemble putative splice variants from overlapping EST and mRNA sequences without resorting to a reference genome, and genome-based methods for clustering spliced alignments and inferring gene models.

Gene indices

Gene indices are gene- or transcript-oriented collections of EST and mRNA sequences grouped by sequence similarity. Traditional methods compare all EST and mRNA sequences against each other to identify significant overlaps, then group and

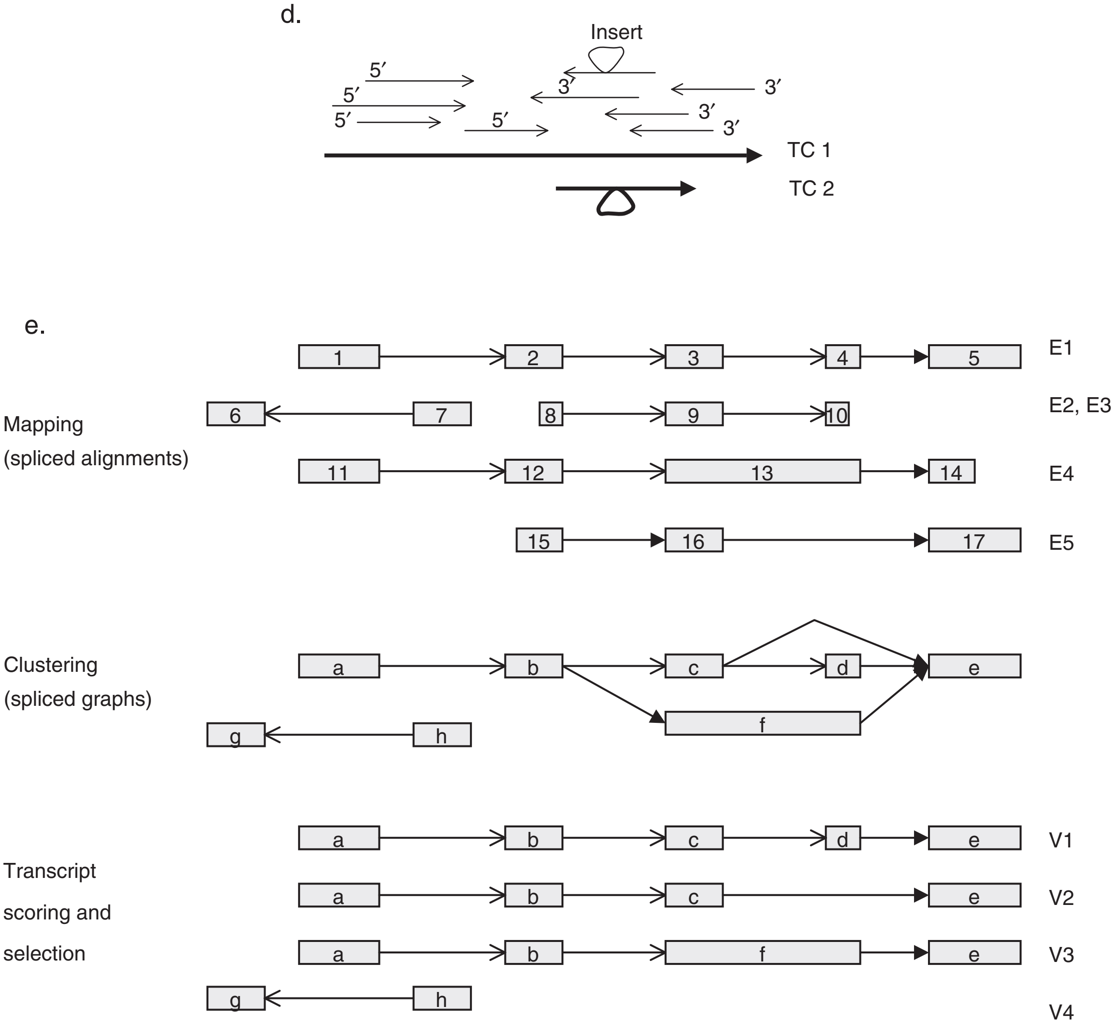

Figure 2: Continued.

assemble sequences with compatible overlaps into disjoint clusters (Figure 2d). The term 'gene indices' may be used to refer to the annotated collections of sequences (UniGene [40]), to the multiple alignments of sequences in the clusters, or to the consensus sequences generated from the multiple alignments (TIGR Gene Indices [41], GeneNest [42]).

Constructing gene indices is complicated by a number of factors. *Over-clustering* may occur when differences between paralogs are interpreted as sequencing errors, forcing paralogs into a same cluster, or when contamination of ESTs with vector or linker sequences or residual polyA tails creates false appearances of overlaps. Other times, insufficient EST sampling may result in *under-clustering*, where several disconnected clusters are produced for a same gene. Furthermore, the process is computationally expensive, as the time required by the pairwise comparisons increases quadratically with the number of sequences in the set. To alleviate these problems, methods were developed to trim the vector and low-quality ends of the EST sequences prior to the comparison, pre-cluster the sequences by aligning EST sequences against the known primary mRNA form (*seeded* clustering), and to *a priori* cluster almost identical or contained sequences to reduce the complexity of the clustering [43].

Genome-based clustering and assembly

In the genome-based approach, the spliced alignments of cDNA and protein sequences are clustered at loci along the genomic sequence. To distinguish among possibly different genes sharing the same genomic locus, including overlapping genes and sense–antisense transcripts [44], clustering is often refined to group alignments separately on each strand of the genomic sequence, and requires that the sequences share common splice junctions in addition to overlapping exons.

One recent innovation in alternative splicing annotation has been the concept of a *splice graph*, as a model for concisely capturing splice variations within a gene. The splice graph represents a gene as a directed acyclic graph in which exons are represented as vertices, introns are the arcs connecting the exons, and splice variants are the paths obtained by traversing the graph from a source vertex (with no incoming arcs) to a sink vertex (with no outgoing arcs). With this set of rules, cDNA and protein spliced alignments at a genomic locus can be consolidated into a spliced graph (Figure 2e) and candidate full-length splice variants can be enumerated from the graph in a combinatorial fashion. The combinatorial nature of the splice graph is also, however, the main limitation of the model, since some of the exon combinations it encodes may be artificial constructs without biological relevance. Different methods have been proposed to select or prioritize candidate transcripts in order to differentiate between those most likely to be biologically relevant and those likely to be artifactual. For instance, the annotation intergrated resource AIR annotation pipeline [37] assigns splice variants confidence scores based on a set of four characteristics measuring the strength of the supporting cDNA and protein evidence, such as the quality and length of the supporting alignments, accuracy of splice signals and the level of fragmentation of the evidence. High-scoring candidates are later selected and promoted into the annotation. The ECgene [45] system classifies candidate transcripts into three categories, high-, medium- and low-confidence, based on the number of cDNA alignments that were stitched together to construct it. Other methods simply defer the task of transcript selection to the users, but provide additional information and visualization to aid in the analysis [46, 47].

The genome-based approaches solve a number of deficiencies observed with the gene indices. By mapping the ESTs to the genome contamination from foreign matter, such as vector and linker sequences, or from polyA tails are removed. In addition, the sequencing errors that were compounded between the cDNA sequences in the gene indices method are now resolved by aligning the sequence against the genome, used as reference. As a result, fewer sequences are misplaced along the genome, resulting in considerably less clustering of paralogs. However, genome-based clustering has its own set of limitations. Some of the remaining issues that have to be resolved involve the contamination of ESTs with genomic fragments or incompletely spliced forms that could produce over-clustering of neighboring or overlapping genes, errors in the alignment, particularly inaccuracy of splice junctions and missing short exons, and strand prediction. Other limitations of the splice graph model include its inability to fully capture splicing variations where one variant is a 5′ or 3′ extension of another, and its sensitivity to alignment errors, as one spurious exon may double the number of candidates.

Identifying splice variation from microarray data

Recently, the Affymetrix chip design using multiple probes per gene has been used to identify splice variations. The Affymetrix GeneChip technology uses 22 probes collected from within the exons or straddling the exon boundaries in the 3′UTR of the sampled genes (Figure 2c). When an exon is alternatively spliced between two microarray experiments, the expression level of its probes in the two arrays will be different and, in at least one of the experiments, significantly dissimilar to those of their neighbors. Thus, statistical tests were developed to determine the significance of these differences and to predict alternative splicing events [48, 49].

RESOURCES FOR ALTERNATIVE SPLICING

The recent years have witnessed a growth in the number of resources—programs, databases and web servers—for the identification and analysis of alternative splicing, directly reflecting the variety of data generation methods. A summary of these resources is given in Table 1.

A growing number of *databases* have emerged that collect instances of alternative splicing events inferred *computationally* from pairwise comparisons of cDNA and protein sequences (PALSdb [52], EASED [55]), their genomic exon-intron structures (ASAP [50], AltExtron and AltSplice [51], SpliceInfo [54]), or GenBank annotations (AsMamDb [53]). In addition to sequence and splicing-related information, they may contain annotations such as tissue specificity, developmental stage, expression data, exon GC and repeat content, conservation in other species and the biological function where known. In contrast, the AEDB database maintained as part of the alternative splicing database (ASD) [51] contains *experimentally* identified alternatively spliced exons and splicing regulatory signals extracted from the literature, with links to relevant MEDLINE entries documenting the events.

In addition to databases, several programs and software pipelines for predicting genes and alternatively spliced transcripts based on genomically aligned cDNA and protein sequences have been developed to support genome annotation projects: ClusterMerge [57] produces the Ensembl ESTgenes, AIR [37] was used to annotate the Celera rat genome, and PASA [56] reconstructed transcripts in the TIGR annotation of *Arabidopsis thaliana*. Outside the spectrum of genome annotation projects, ASG [46] provides internet accessible splice graphs for several eukaryotic genomes and ASmodeler [58] allows users to create their own splice graph annotations via a web interface.

Lastly, the most comprehensive gene index collections to date are the TIGR Gene Indices [41] and the NCBI UniGene collection [40].

Table 1: Alternative splicing annotation resources in the literature and on the World Wide Web

Resource	Type of data	Method	Address	Dist.*
Alternative splicing events and partial splice forms				
ASAP [50]	Alternative splice forms (events)	Comparison of cDNA–genomic alignments	http://www.bioinformatics.ucla.edu/ASAP	DB
ASD [51] (AltExtron, AltSplice)	Alternative splicing events	Comparison of cDNA–genomic alignments	http://www.ebi.ac.uk/asd	DB
ASD [51] (AEDB)	Alternatively spliced exons	Experimental data collected from the literature	http://www.ebi.ac.uk/asd	DB
PALSdb [52]	Alternative splice forms (events)	Comparison of mRNA and EST sequences from UniGene clusters	http://palsdb.ym.edu.tw/	DB
AsMamDB [53]	Alternative splicing events	Comparison of GenBank annotated splice variants	http://166.111.30.65/ASMAMDB/	DB
SpliceInfo [54]	Alternative splicing events	Comparison of mRNA–and protein–genomic alignments	http://SpliceInfo.mbc.NCTU.edu.tw	DB
EASED [55]	Alternative splice forms (events)	Comparison of ESTs with mRNAs	http://eased.bioinf.mdc-berlin.de/	DB
Prediction of alternatively spliced transcripts				
TAP [4]	Prediction of primary transcript, other splice forms	Genome-based primary transcript prediction from maximal splice junction chains	http://sapiens.wustl.edu/~zkan/TAP	SC; WS
PASA [56]	Transcript prediction and selection	Genome-based transcript prediction and selection	http://www.tigr.org/tdb/e2k1/ath1/pasa.annot.updates/pasa.annot.updates.shtml	DB; SC
AIR [37]	Gene and transcript prediction and selection	Genome-based splice graph for transcript prediction and selection	https://panther.appliedbiosystems.com/publications.jsp	SC; DB
ClusterMerge [57]	Maximal EST- assembled transcripts	Genome-based assembly of compatible alignments	http://www.ensembl.org	DB; SC
ASmodeler [58], ECgene [45]	Gene and transcript prediction	Genome-based splice graphs for transcript prediction and selection	http://genome.ewha.ac.kr/ECgene/ASmodeler	WS; DB
ASG [46]	Splice graphs, combinatorial splice variants	Genome-based splice graph enumeration of transcripts	http://statgen.ncsu.edu/asg/	WS; DB
TIGR [41]	Gene indices (consensus transcripts)	Off-genome cDNA clustering and assembly	http://www.tigr.org/tdb/tgi/	DB
UniGene [40]	Gene indices (gene clusters)	Off-genome & genome-based cDNA clustering	http://www.ncbi.nlm.nih.gov	DB
GeneNest [42]	Gene indices (consensus transcripts)	Off-genome cDNA clustering and assembly	http://genenest.molgen.mpg.de	DB

*Type of distribution: DB = database or data set; WS = web server; SC = source code

HURDLES IN ALTERNATIVE SPLICING ANNOTATION

Although tremendous strides were taken in the past few years to identify and catalogue splice variation in the human and other eukaryotic genomes, this chapter of genomics is still very much a draft. Among the factors that continue to confound annotation methods are:

(i) EST quality—contamination of EST sequences with genomic, cross-species and vector sequences, high sequencing error rates, chimeras;
(ii) alignment accuracy—ambiguity of predicted strands, alignment errors, particularly at the splice junctions, missing exons;
(iii) limitations in models and data—insufficient data, for instance for tissue-specific clustering, EST fragmentation; and
(iv) complexity of biological processes—paralogs, trans-splicing, sense–antisense transcripts and overlapping genes.

With its far-reaching implications for biology and medicine, identifying and characterizing alternative splicing variation is becoming an important part of the effort for detecting and cataloguing the genes. Thus, developing a comprehensive catalogue of splicing variation will require mobilization to produce fast and accurate tools for automatic annotation, coupled with an assiduous curation effort and high-throughput validation by sequencing, RT–PCR and microarray experiments [59].

ALTERNATIVE SPLICING REGULATION

An equally intense area of research in alternative splicing revolves about the factors and mechanisms of splicing regulation.

Pre-mRNA splicing is a complex cellular process undertaken in the cytoplasm by the spliceosome. The spliceosome assembles onto each intron from a set of five small nuclear ribonucleoproteins (snRNPs U1, U2, U4–U6) and numerous accessory proteins that bind specifically to locations at or within the vicinity of the splice sites, and catalyzes the excision of the intron [60]. While the 5′ (donor) and 3′ (acceptor) splice sites have well characterized consensus sequences that are recognized to play a major role in splicing, an increasing body of evidence reveals that previously unknown RNA elements located outside the splice signals, in exons and introns, contribute to the exon's inclusion or exclusion in the mature mRNA, in a network of interactions that appear to be centered on exons, rather than introns. These *cis*-regulatory elements can promote (*splicing enhancers*) or repress (*splicing silencers*) the inclusion of the exon in the mRNA through the activity of the bound regulatory proteins, and can be located in the exons—*exonic splicing enhancers (ESEs) and silencers (ESS)*, or introns—*intronic splicing enhancers (ISE) and silencers (ISS)*. They can act from both within the proximity of the exon, or from 300–1000 bp away. It is becoming increasingly evident that many exons, constitutive or alternative, and their surrounding introns harbor both silencing and enhancing elements, and that the exon's inclusion/exclusion is the result of competition between the two effects [61].

Mechanisms of splicing activation

Exonic enhancers are the binding sites of splicing activator proteins, the most studied activators being the SR (Ser/Arg) factors. The SR proteins are characterized by the presence of 1–2 RNA recognition motifs (RRM), and a C-terminal RS-domain enriched in Arg/Ser dipeptides. The primary mechanism by which they promote splicing is by attaching to enhancer elements within the exon or in the polypyrimidine tract via their RRM domain and recruiting the U2AF factor with their RS domain. An RS-domain independent mechanism has also been hypothesized, in which the splicing factor binds to an ESE to antagonize the effect of a neighboring silencer [62] (Figure 3a,b). Exonic splicing enhancer motifs have been characterized through a variety of bioinformatics methods and through functional experiments. Their mechanism, diversity and action were reviewed in [63].

Mechanisms of splicing repression

Splicing silencers are thought to be binding sites of splicing repressor proteins. Splicing repression is effected largely through the activity of the hnRNP protein families. Of these, hnRNP I, also known as PTB [64], members of the hnRNP A/B families, and the neuron-specific NOVA-1 inhibitor [65] are among the best characterized. Several mechanisms have been proposed for hnRNP-mediated silencing (Figure 3c–e): by interfering with the spliceosome

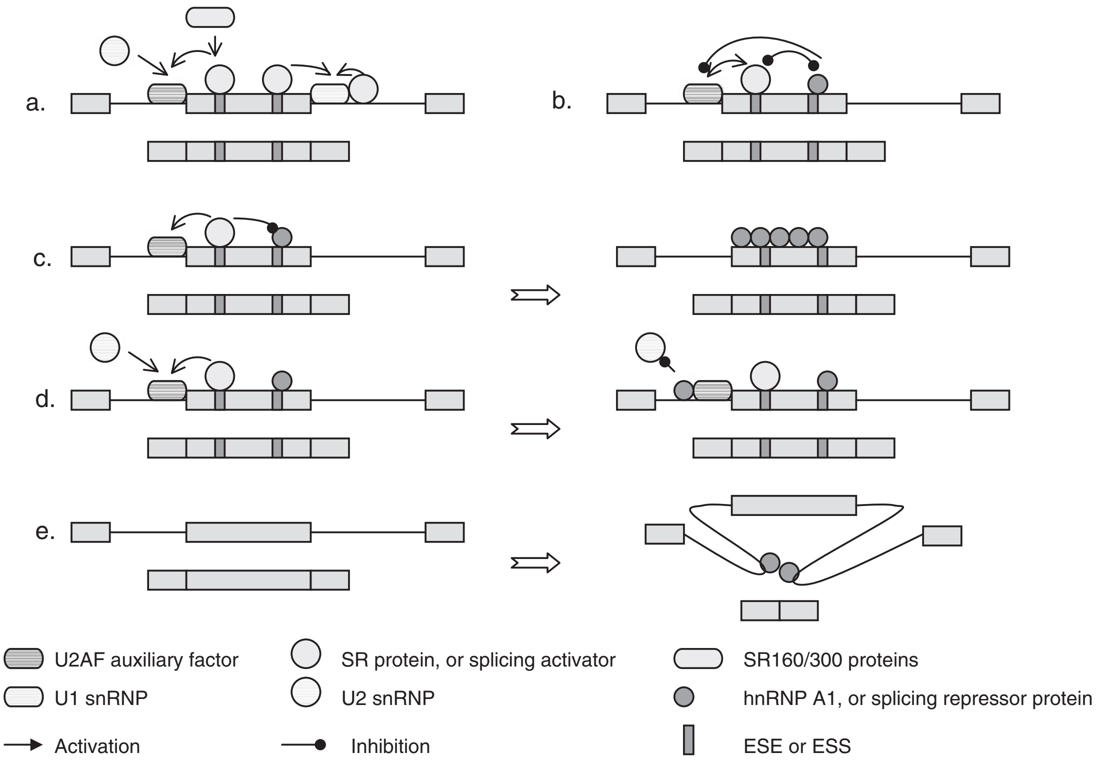

Figure 3: Mechanisms of alternative splicing regulation: (**a**, **b**). activation; (**c**–**e**). repression. (**a**) Splicing *activation* is enabled primarily through the activator function of the RS-domain of an SR protein bound to an ESE element through its RRM domain, (**b**) or can be RS-domain independent, with the activator blocking the action of a nearby silencing element in direct competition. (**c**) Splicing *repression* is accomplished primarily by hnRNP proteins, either by neutralizing the effect of an activator protein through direct competition, (**d**) or by cooperative binding of several silencing factors that displace the ESE-bound activators (**e**), or by binding to motifs on either side of the exon followed by dimerization, which places the exon in a loop inaccessible to the splicing apparatus.

assembly through cooperative binding of several inhibitory elements, by blocking neighboring ESEs, or by binding to duplicate intronic sequences on both sides of the exon to form a loop that renders the exon inaccessible to the spliceosome. Experimentally identified or validated ISS sequences were reviewed in [66].

BIOINFORMATICS METHODS FOR IDENTIFYING SPLICING REGULATORY ELEMENTS

Methods for identifying and characterizing splicing regulatory elements are still in their infancy, varying from experimental to computational and from focused to large-scale, and can be centered on the splicing factors (proteins) or their binding sites. Emphasizing that all types of approaches have produced valuable results, we hereto focus on computational methods. While there are many similarities with the detection of transcriptional regulatory sites, the specific challenges associated with identifying splicing regulatory sites arise from the fact that few proteins have been identified, and that the full extent of splicing regulatory factors, including the number of proteins and the prevalence of their binding sites on the pre-RNA sequence, cannot be estimated yet.

Consensus sequences

Most early work on splicing enhancers targeted purine-rich exonic elements that were thought to be binding sites of SR proteins (Table 2) and produced a number of site consensus sequences [62]. A consensus sequence uses IUPAC codes to specify what symbols are allowed at every position in the motif's sequence, but cannot quantify the preference for certain symbols at a given motif position. The purine-rich model for regulatory sequences was put to rest when ESE sequences in IgM and in the avian sarcoma pre-mRNA were found to retain their

Table 2: Consensus motifs for SR protein dependent enhancers [62]. Reproduced with permission from *Nature Reviews Genetics*, Vol. 3, pp. 385–398, copyright (2002) Macmillan Magazines Ltd

Protein	High-affinity binding site*	Functional ESE*
SRp20	WCWWC	GCUCCUCUUCC
	CUCKUCY	CCUCGUCC
SC35	AGSAGAGUA	GRYYMCYR
	GUUCGAGUA	UGCYGYY
	UGUUCSAGUAGKS	
	AGGAGAU	
9G8	(GAC)n	
	ACGAGAGAY	
	WGGACRA	
SF2/ASF	RGAAGAAC	CRSMSGW
	AGGACRRAGC	
SRp40	UGGGAGCRGUYRGCUCGY	YRCRKM
SRp55		YYWCWSG
TRA2β	(GAA)n	

*IUPAC codes: M = A or C; R = A or G; W = A or U; Y = C or U; S = C or G; K = G or U

effect *in vitro* even after their purine-rich stretches were eliminated [67, 68].

Position weight matrices

Position weight matrices (PWMs) are a more suitable model for specifying the variability of symbols in a motif sequence. The PWMs are matrices of scores constructed from the nucleotide frequencies at each position within the motif, typically derived from a collection of experimentally identified sites (Figure 4). A PWM specifies a score for each nucleotide at every motif position. The score of a candidate oligonucleotide sequence is then computed as the sum of the individual nucleotide score contributions, and used as an indicator for the reliability of the match: the higher the score, the stronger and more reliable the match. The PWMs for four splicing activator factors, SF2/ASF, SC35, SRp40 and SRp55, were constructed from alignments of binding sites identified with functional *in vivo* and *in vitro* systematic evolution of ligands by exponential (SELEX) enrichment experiments using a minigene construct to test the effects of ESEs on splicing [69]. These are the only experimentally derived PWMs of splicing factors to date.

Statistical over-representation of *k*-mers

Statistical methods, emblematic for post-genome large-scale genomics, aim at identifying sequence elements that are enriched in one class of contexts, defining the biological setting of the element sought, compared with another, defining the background or null setting. Typically, the candidate sequence elements are simple oligonucleotides of length *k* (*k*-mers), thought to form the core of splicing regulatory motifs. The methods start with the set of all *k*-mers, for a value of *k* selected based on statistical or empirical considerations, and assign them scores that measure the difference between the oligonucleotide frequencies in the two types of contexts. Thus, the first step of such methods consists of identifying a number of attributes that biologically characterize the elements, and the two sets of contexts, 'functional' versus 'null', for each attribute. Each candidate *k*-mer is then assigned a value for each attribute that statistically characterizes its relative positioning between the two contexts. Lastly, a set of candidate *k*-mers that show statistically significant differences between the two contexts for each of the attributes is selected for further analysis or for validation *in vivo* or *in vitro*.

Several variations of this model have been used under different biological assumptions, or attributes. Fairbrother *et al.* [71] started from the following assumptions about ESE elements: (i) ESE motifs are more likely to occur in exons than in introns; and (ii) ESE motifs are more likely to occur in weak exons (with non-consensus splice sites) than strong exons (with consensus splice sites). For $k = 6$, the score function for each of the 4096 6-mers was calculated as the difference between the frequency of the *k*-mer in exons versus introns, and in weak versus strong exons, scaled by the standard deviation of the sample. Candidates that showed significant enrichment in both tests, corresponding to a scaled difference value larger than 2.5 (or equivalently, with scaled differences >2.5 standard deviations) were selected. These hexamer sequences were then clustered by sequence similarity and aligned to form PWMs for ESE motifs. A similar approach led to the computational rediscovery of the UGCAUG hexanucleotide motif in introns proximal to brain-tissue specific alternative exons, by comparing the frequency of the hexamer in introns downstream of constitutive versus alternatively spliced exons [72]. The motif had been previously experimentally identified as part of a splicing enhancer for exon N1 in the *c-src* gene [73]. In another study, Zhang and Chasin [74] focused on constitutive splicing and searched for 8-mers over-represented in internal noncoding exons, thought to contain signals a.

Alignment

```
ATACCTTAT
TTACTTAAT
TTACAATAT
ATACTAAAT
ATGCTTTAT
```

Consensus

WTRCNWAT

W = A or T
R = purine (A or G)
N = any nucleotide (A, C, G, T)

b.

Alignment

```
ATACCTTAT
TTACTTAAT
TTACAATAT
GTACTACAT
ATACTAAAT
ATGCTTTAT
```

Count matrix

N	1	2	3	4	5	6	7	8	9
A	3	0	5	0	1	3	2	6	0
C	0	0	0	6	1	0	1	0	0
G	1	0	1	0	0	0	0	0	0
T	2	6	0	0	4	3	3	0	6

Position weight matrix (PWM)

$W(a,i) = \log (f_{a,i} / p_a)$,
$i=1..9$, $a = A, C, G$ or T

$f_{a,i} = N(_{a,i}) / \Sigma_{b=A,C,G,}$ TN(b,i)

p_a= background probability of 'a'

... TTACT**ATACAACGT**AGCTAGCATC ...

Score(**ATACAACGT**) = W(A,1)+W(T,2)+ ... +W(G,8)+W(T,9)

Figure 4: Bioinformatics representations of splicing regulatory motifs: (**a**) consensus sequence and (**b**) position weight matrix (PWM). A consensus sequence specifies the set of bases that can appear at each position in the motif, using accepted IUPAC letter codes to represent combinations of bases. A PWM is a more expressive representation, incorporating the expected proportions of bases at each position in the motif. A PWM has a score for each base (A, C, G, T shown on rows) at each position in the motif sequence (shown along columns), derived from the site nucleotide counts (*N*) and frequencies (*f*) with a log-odds formula [70]. The score for a candidate nucleotide sequence is the sum of individual nucleotide scores at the corresponding motif positions: the higher the score, the better the match and the more likely to be real.

that would mediate splicing, compared to unspliced pseudoexons and 5′ UTRs of transcripts of intronless genes.

Motif discovery

Motif discovery methods commonly used in identifying transcription factor binding motifs in unaligned sequences, such as MEME [75] and Gibbs sampling [76], can also be employed to identify splicing regulatory signals. Although they achieved some success in identifying two regulatory motifs in introns surrounding 54 alternatively spliced exons [77], it is hard to estimate what their predictive power will be in a large set of diverse sequences, harboring a potentially wide variety of regulatory signals with low sequence specificity. For best results, pre-selecting and curating the data, for instance by selecting tissue-specific events, or validating the results with an independent method, such as comparative studies of alternatively spliced exons in the human and mouse species [77], is essential.

Comparative studies

Conserved regions in alignments of orthologous *introns* of human and mouse, or similarly related species, provide good candidates for functional sequences. Surprisingly, in the context of alternative splicing, comparative methods have focused less on predicting specific splicing regulatory signals and more on characterizing the large-scale patterns of sequence conservation [78, 47], or on validating motifs identified by other methods [77]. For instance, intronic sequences flanking alternatively spliced exons compared to constitutive exons were found to be more strongly conserved between human and mouse, thus suggesting that their splicing mechanisms may be different from those of constitutively spliced exons [78].

Identification of regulatory signals within coding sequences of *exons* by sequence comparison is further complicated by the fact that additional selective pressure is imposed on the coding sequence by the necessity to preserve the protein. Several studies showed that purifying selective pressure may act against silent mutations in certain regions in the coding sequence [79], and that the phenomenon is more pronounced in alternatively spliced exons [80]. There are also indications that exonic portions of splice sites may undergo positive selection to stabilize during the evolution of introns after insertion [81]. These findings may herald new methods, based on sequence evolutionary models and tools and taking advantage of the large volumes of multi-species data becoming available. Already, Blanchette [82] used phylogenetic methods to distinguish between conservation due to amino acid constraints and conservation due to regulatory factors in predicting splicing regulatory signals in coding sequences.

RESOURCES FOR ALTERNATIVE SPLICING REGULATION

The repertoire of resources to aid in annotating splicing regulatory regions reflects the early state of bioinformatics and experimental studies in this area. A web-based tool, called ESEfinder, was implemented to allow users to search their sequences for matches to the binding motifs of four splicing activators identified by SELEX experiments, SF2/ASF, SC35, SRp40, SRp55 (http://rulai.cshl.edu/tools/ESE/) [83]. Another web-based tool, RESCUE-ESE [84], searches user-input sequences for the presence of specific 6-mer motifs previously identified as potential cores of splicing regulatory motifs by statistical over-representation methods (http://genes.mit.edu/burgelab/rescue-ese/). The SpliceInfo [54] alternative splicing environment (http://spliceinfo.mbc.nctu.edu.tw/) provides tools for the discovery of sequence or structural motifs (Mfold [85], Gibbs sampling [76], MEME[75]) in sequences of annotated exons and introns, while the AEDB [51] database (http://www.ebi.ac.uk/asd/aedb/) collects experimentally identified splicing regulatory sites and sequences from the literature. Lastly, more general sequence alignment and visualization tools such as PipMaker [86] (http://pipmaker.bx.psu.edu/pipmaker/) and Vista [87] (http://genome.lbl.gov/vista/mvista/submit.shtml) can aid in the discovery of intronic regulatory signals.

CONCLUSIONS

Alternative splicing is without doubt one of the most important gene regulatory mechanisms, and one that has reemerged as a central concept in the post-genome sequencing era. Determining the extent and importance of alternative splicing required the confluence of critical advances in data acquisition, improved understanding of biological processes and the development of fast and accurate computational analysis tools.

While tremendous progress has been made in annotating alternative splicing variations and exploratory steps were taken into the regulation of alternative splicing, there is a long path to fully understanding how it manifests itself in the cell and what are the determining factors to be able to use and interpret this information in disease diagnostics and treatment. In creating an alternative splicing compendium, comparative and functional genomics efforts must address immediate challenges such as improving the accuracy of annotations by differentiating informative events from experimental or computational artifacts and by evaluating the accuracy of existing methods, as well as medium- to long-term goals, including the systematic generation of EST and mRNA sequences from a variety of tissues and under various cell conditions, the development of new methods for the prediction and analysis of splicing regulatory elements and mechanisms, and breakthroughs in high-throughput technologies for *in vitro* and *in vivo* validation and functional characterizations. These are the first steps toward the genetic medicine of the future.

Key Points

- Alternative splicing has important implications for biology and medicine.
- Identifying all splice variations and the signals that regulate splicing are core problems in genomics and bioinformatics.
- Large EST data sets and sequence data from multiple genomes are fueling efforts to develop new computational methods to address these problems.

Acknowledgment

Thanks to Rosane Charlab for her careful review of the manuscript.

References

1. Gilbert W. Why genes in pieces?. *Nature* 1978;**271**:501.
2. Breitbart RE, Andreadis A, Nadal Ginard B. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Ann Rev Bioch* 1987;**56**:467–95.
3. Mironov AA, Fickett JW, Gelfand MS. Frequent alternative splicing of human genes. *Genome Res* 1999;**9**:1288–93.

4. Kan Z, Rouchka EC, Gish WR, *et al*. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res* 2001;**11**:889–900.
5. Modrek B, Resch A, Grasso C, *et al*. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucl Acids Res* 2001;**29**:2850–59.
6. Zavolan M, Kondo S, Schönbach C, *et al*. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res* 2003;**13**:1290–300.
7. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 2003;**428**:493–21.
8. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004;**431**:931–45.
9. Graveley BR. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* 2001;**17**:100–7.
10. Carstens RP, Wagner EJ, Garcia-Blanco MA. An intronic splicing silencer causes skipping of the IIIb exon of fibroblast growth factor receptor 2 through involvement of poly-pyrimidine tract binding protein. *Mol Cell Biol* 2000;**20**: 7388–400.
11. Saxena S, Szabo CI, Chopin S, *et al*. BRCA1 and BRCA2 in Indian breast cancer patients. *Hum Mutat* 2002; **20**:473–74.
12. Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. *Genes Dev* 2000;**17**:419–37.
13. Garcia-Blanco MA, Baraniak AP, Lasda EL. Alternative splicing in disease and therapy. *Nat Biotechnol* 2004;**22**: 535–46.
14. Muraki M, Ohkawara B, Hosoya T, *et al*. Manipulation of alternative splicing by a newly developed inhibitor of Clks. *J Biol Chem* 2004;**279**:24246–54.
15. Cartegni L, Krainer AR. Correction of disease-associated exon skipping by synthetic exon-specific activators. *Nat Struct Biol* 2003;**10**:120–25.
16. Lopez AJ. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu Rev Genet* 1998;**32**:279–305.
17. Lorson CL, Hahnen E, Androphy EJ, *et al*. A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc Natl Acad Sci USA* 1999;**96**:6307–11.
18. Lee C, Wang Q. Bioinformatics analysis of alternative splicing. *Brief Bioinform* 2005;**6**:23–33.
19. Zhang MQ. Computational prediction of eukaryotic protein-coding genes. *Nat Reviews Genet* 2002;**3**:698–709.
20. Hubbard T, Barker D, Birney E, *et al*. The Ensembl genome database project. *Nucleic Acids Res* 2002;**30**:38–41.
21. Karolchik D, Baertsch R, Diekhans M, *et al*. The UCSC Genome Browser Database. *Nucleic Acids Res* 2003;**31**: 51–54.
22. Venter JC, Adams MD, Myers EW, *et al*. The sequence of the human genome. *Science* 2001;**291**:1304–51.
23. Stamm S, Ben-Ari S, Rafalska I, *et al*. Function of alternative splicing. *Gene* 2005;**344**:1–20.
24. Stamm S, Zhu J, Nakai K, *et al*. An alternative-exon database and its statistical analysis. *DNA Cell Biol* 2000;**19**: 739–56.
25. Akker SA, Smith PJ, Chew SL. Nuclear post-transcriptional control of gene expression. *J Mol Endocrinol* 2001;**27**:123–31.
26. Fehlbaum P, Guihal C, Bracco L, *et al*. A microarray configuration to quantify expression levels and relative abundance of splice variants. *Nucl Acids Res* 2005;**33**:e47.
27. Hide WA, Babenko VN, van Heusden PA, *et al*. The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res* 2001;**11**:1848–53.
28. Boguski MS, Lowe TM, Tolstoshev CM. dbEST—database for expressed sequence tags. *Nat Genet* 1993;**4**:332–33.
29. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence project: update and current status. *Nucl Acids Res* 2003;**31**:34–37.
30. Strausberg RL, Feingold EA, Klausner RD, *et al*. The mammalian gene collection. *Science* 1999;**286**:455–57.
31. Bairoch A, Apweiler R, Wu CH, *et al*. The Universal Protein Resource (UniProt). *Nucleic Acids Res* 2005;**33**: D154–59.
32. Mott R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput Appl Biosci* 1997;**3**:477–78.
33. Florea L, Hartzell G, Zhang Z, *et al*. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* 1998;**8**:967–74.
34. Wheelan SJ, Church DM, Ostell JM. Spidey: a tool for mRNA-to-genomic alignments. *Genome Res* 2001;**11**: 1952–57.
35. Usuka J, Zhu W, Brendel V. Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics* 2000;**16**:203–11.
36. Kent WJ. BLAT–the BLAST-like alignment tool. *Genome Res* 2002;**12**:656–64.
37. Florea L, Di Francesco V, Miller J, *et al*. Gene and alternative splicing annotation with AIR. *Genome Res* 2005; **15**:54–66.
38. Lee BT, Tan TW, Ranganathan S. MGAlignIt: a web service for the alignment of mRNA/EST and genomic sequences. *Nucleic Acids Res* 2003;**31**:3533–36.
39. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 2005;**21**:59–75.
40. Schuler GD. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med* 1997;**75**:694–98.
41. Liang F, Holt I, Pertea G, *et al*. An optimized protocol for analysis of EST sequences. *Nucl. Acids Res* 2000;**28**: 2657–65.
42. Haas SA, Beissbarth T, Rivals E, *et al*. GeneNest: automated generation and visualization of gene indices. *Trends Genet* 2000;**16**:521–23.
43. Lee Y, Tsai J, Sunkara S, *et al*. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res* 2005;**33**(Database issue):D71–4.
44. Kiyosawa H, Mise N, Iwase S, *et al*. Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. *Genome Res* 2005; **15**:463–74.
45. Kim N, Shin S, Lee S. ECgene: genome-based EST clustering and gene modeling for alternative splicing. *Genome Res* 2005;**15**:566–76.

46. Leipzig J, Pevzner P, Heber S. The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Res* 2004;**32**:3977–83.

47. Sugnet CW, Kent WJ, Ares Jr M, *et al*. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac Symp Biocomput* 2004;**9**:66–77.

48. Hu GK, Madore SJ, Moldover B, *et al*. Predicting splice variant from DNA gene expression data. *Genome Res* 2001; **11**:1237–45.

49. Wang H, Hubbell E, Hu J, *et al*. Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics* 2003;**19**(Suppl.1):i315–22.

50. Lee C, Atanelov L, Modrek B, *et al*. ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res* 2003;**31**: 101–5.

51. Thanaraj TA, Stamm S, Clark F, *et al*. ASD: the Alternative Splicing Database. *Nucleic Acids Res* 2004;**32**(Database issue): D64–69.

52. Huang YH, Chen Y.T, Lai JJ, *et al*. PALS db: Putative Alternative Splicing database. *Nucl Acids Res* 2002; **30**:186–90.

53. Ji H, Zhou Q, Wen F, *et al*. AsMamDB: an alternative splice database of mammals. *Nucleic Acids Res* 2001;**29**: 260–63.

54. Huang HD, Horng JT, Lin FM, *et al*. SpliceInfo: an information repository for mRNA alternative splicing in human genome. *Nucleic Acids Res* 2005;**33**(Database issue): D80–85.

55. Pospisil H, Herrmann A, Bortfieldt RH, *et al*. EASED: Extended Alternatively Spliced EST Database. *Nucl Acids Res* 2004;**32**:D70–74.

56. Haas BJ, Delcher AL, Mount SM, *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 2003;**31**:5654–66.

57. Eyras E, Caccamo M, Curwen V, *et al*. ESTGenes: alternative splicing from ESTs in Ensembl. *Genome Res* 2004;**14**:976–87.

58. Kim N, Shin S, Lee S. ASmodeler: gene modeling of alternative splicing from genomic alignment of mRNA, EST and protein sequences. *Nucleic Acids Res* 2004;**32**(Web Server issue):W181–86.

59. Johnson JM, Castle J, Garrett-Engele P, *et al*. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 2003;**302**:2141–44.

60. Black DL. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 2003;**72**:291–36.

61. Smith C, Valcarcel J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci* 2000;**25**: 381–88.

62. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* 2002;**3**:285–98.

63. Blencowe BJ. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* 2000;**25**:106–10.

64. Wagner EJ, Garcia-Blanco MA. Polypyrimidine tract binding protein antagonizes exon definition. *Mol Cell Biol* 2001;**21**:3281–88.

65. Jensen KB, Dredge BK, Stefani G, *et al*. Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron* 2001;**25**:359–71.

66. Ladd AN, Cooper TA. Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol* 2002;**3**(reviews):0008.1–0008.16.

67. Watakabe A, Tanaka K, Shimura Y. The role of exon sequences in splice site selection. *Genes Dev* 1993;**7**: 407–18.

68. Staknis D, Reed R. SR proteins promote the first specific recognition of pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex. *Mol Cell Biol* 1994;**14**: 7670–82.

69. Liu HX, Zhang M, Krainer AR. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* 1998;**12**:1998–12.

70. Stormo G. DNA binding sites: representation and discovery. *Bioinformatics* 2000;**16**:16–23.

71. Fairbrother WG, Yeh RF, Sharp PA, *et al*. Predictive identification of exonic splicing enhancers in human genes. *Science* 2002;**297**:1007–13.

72. Brudno M, Gelfand MS, Spengler S, *et al*. Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucl Acids Res* 2001;**29**:2338–48.

73. Modafferi EF, Black DL. A complex intronic splicing enhancer from the *c-src* pre-mRNA activates inclusion of a heterologous exon. *Mol. Cell. Biol* 1997;**17**: 6537–45.

74. Zhang XH, Chasin LA. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* 2004;**18**:1241–50.

75. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc 2nd Intl Conf on Intell Sys Mol Biol* 1994;28–36.

76. Lawrence CE, Altschul SF, Boguski MS, *et al*. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993;**262**:208–14.

77. Miriami E, Margalit H, Sperling R. Conserved sequence elements associated with exon skipping. *Nucl Acids Res* 2003;**31** 1974–83.

78. Sorek R, Ast G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res* 2003;**13**:1631–37.

79. Hurst LD, Pal C. Evidence for purifying selection acting on silent sites in BRCA1. *Trends Genet* 2001;**17**:62–65.

80. Iida K, Akashi H. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene* 2000;**261**: 93–105.

81. Sverdlov AV, Rogozin IB, Babenko VN, *et al*. Evidence of splice signal migration from exon to intron during intron evolution. *Curr Biol* 2003;**13**:2170–74.

82. Blanchette M. A comparative analysis method for detecting binding sites in coding regions. *Proc 7th Annual Intl Conf on Comp Biol - RECOMB 2003* 2003;57–66.

83. Cartegni L, Wang J, Zhu Z, *et al*. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucl. Acids Res* 2003;**31**:3568–71.

84. Fairbrother WG, Yeo GW, Yeh R, *et al*. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucl Acids Res* 2004;**32**(Web Server issue):W187–90.

85. Zuker M, Mathews DH, Turner DH. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. *RNA Biochemistry and Biotechnology*. Barciszewski J and Clark BFC (eds), NATO ASI Series: Kluwer Academic Publishers, 1999.

86. Schwartz S, Zhang Z, Frazer KA, *et al*. PipMaker–a web server for aligning two genomic DNA sequences. *Genome Res* 2000;**10**:577–86.

87. Frazer KA, Pachter L, Poliakov A, *et al*. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 2004;**32**(Web Server issue):W273–79.