Published in final edited form as:
Nat Biotechnol. 2010 July 1; 28(7): 686–687. doi:10.1038/nbt0710-686.

Splicing by cell type

Mauricio A. Arias, **Shengdong Ke**, and **Lawrence A. Chasin**
Department of Biological Sciences, Columbia University, New York, NY USA.

The rules governing the inclusion of alternative exons in different cell types to generate protein diversity are complex and apparently manifold. In a recent paper in Nature, Barash et al.1 have applied machine learning to high throughput splicing data to identify combinations of sequence features that can be used to predict tissue-specific alternative splicing patterns.

In most eukaryotes the central dogma for the flow of information from gene to protein is: DNA makes pre-mRNA makes mRNA makes protein. In the first step, transcription, genetic information is mapped from DNA to RNA via a one-to-one correspondence between like nucleotides. In the biochemically complex third step, translation, a simple code is used to place amino acids into polypeptides according to sequentially read base triplets in mature messenger RNA. In the middle step, RNA processing, the retrieval of genetic information is less straightforward. The parts of the gene sequence in the primary transcript that are destined to code for protein (the exons) are extracted from a much longer sequence and spliced together, creating a messenger RNA and discarding the intervening sequences (the introns) in the process. This extraction requires, first and foremost, the identification of the exons and the introns. Intron recognition is always taking place during the splicing event itself, a chemical reaction catalyzed by the spliceosome, which is a large molecular machine comprised of 5 RNA molecules and over 100 proteins. On the other hand, an early step in splicing is thought to be the definition of an exon as an entity unto itself, the main evidence for the latter being that disruption of an individual splice site most often leads to the entire exon being discarded, i.e., skipped.

How this early exon recognition takes place is not well-understood. There is not enough information in the splice site sequences themselves to demarcate their borders, and several types of experiments have shown that additional information exists in short degenerate sequence motifs that lie both within and outside the exons. It is thought and has often been demonstrated that these genetic elements interact with specific RNA-binding proteins to either enhance or silence splicing, but the mechanism(s) by which this enhancement or silencing is realized has remained elusive. The nature and location of these sequence elements has been termed the “splicing code”2–5. Reading this code is probably more

Corresponding authors Correspondence to: Lawrence Chasin (lac2@columbia.edu).
All three contributed equally to this article.

Competing financial interests
The authors declare no competing financial interests.

complicated than looking at the linear arrangement of these sequence elements, for at least two reasons. First, RNA can fold into intricate 3-dimensional structures, driven mostly by base pairing between different regions of the molecule. Thus the availability of an RNA sequence to bind an RNA binding protein is modulated by RNA structure, itself driven by an RNA structure code. Pre-mRNA structure *per se* could also play a direct role in splicing. Second, since splicing can take place while RNA is being transcribed, it can be influenced by the transcription complex, which may act as a conduit for the delivery of gene-specific splicing factors and/or by pausing transcription to allow a splice site to be recognized6. In an analogous way, chromatin structure is emerging as a possible modulating factor in code reading (e.g.,7, 8). Thus the splicing code can be comprised of sequences that act at the level of DNA as well as RNA.

It gets more complicated, because: 1) the splicing code does not always produce an all-or-nothing result; and 2) the code can be interpreted differently in different cellular environments. The result is alternative splicing, with the same gene giving rise to multiple mRNA isoforms with different exon constituents and their attendant protein isoforms. Although most exons are spliced constitutively, i.e., included with near 100% efficiency in all mature mRNA molecules produced in all tissues, a large minority are alternatively spliced, such that almost all mammalian genes undergo some alternative splicing. It is hard to overestimate the importance of alternative splicing, as it can generate a proteome that is much greater than the transcriptome, explaining the relative complexity of higher organisms without much of a difference in genome size. Indeed, most research on splicing has focused on this phenomenon. Tissue-specific alternative splicing adds another layer to the splicing code, sequences that allow an exon to be included (or included more often) in one tissue compared to another, with the different behaviors presumably being mediated by different repertoires or levels of splicing factors and/or chromatin structures. This tissue-specific alternative splicing code may be part and parcel of the general code, or distinct from it, or the two may overlap.

How can we go about revealing such a specialized complex code that may depend on multiple variables, some of which are likely unknown as yet. A recent paper by Barash et al. 1 provides one impressive answer. This work was the fruit of an ongoing collaboration between the computational biology lab of Brendan Frey and the splicing lab of Ben Blencowe, both at the University of Toronto. The strategy (Fig. 1) was to reveal the elements of the tissue-specific splicing code by associating the presence of sequence “features” (more later) with the differential inclusion of alternatively spliced exons. The latter was determined using high-throughput microarray measurements of mRNA levels comprising 3665 alternatively spliced exons in 27 mouse cells and tissues. The complexity of the problem was then reduced in two ways. 1) The 27 samples were grouped into four tissue categories for comparison (those related to the CNS, muscle, digestion, and the embryo). 2) Relative %inclusion levels were discretized as three probabilities: increased, decreased or unchanged inclusion in a particular tissue compared to a baseline. Next, a machine learning algorithm was developed to discover which features were associated with increased or decreased exon inclusion in each tissue category. The algorithm was tested against exons not used for training for its ability to predict increased or decreased relative inclusion levels in pairwise comparisons of different tissue categories. An accuracy of about 90% was achieved, attesting to the validity of the method.

Now to the features, perhaps the heart of this work. Barash et al. compiled a list of 1014 diverse features based on previous studies and on their own intuition. Most were based on oligomeric sequences from different types of experiments: e.g., sets of predicted and validated hexamer sequences based on statistical analysis of the transcriptome, others based on evolutionary conservation, ligand sequences for splicing factors, and positional weight matrices for sequences derived by functional selection. But then the sequence list went on to include the density of all possible base trimers, dimers, and even single bases. RNA structure was considered as predicted singlestrandedness around regions such as the splice sites. Splice site scores, the creation of premature stop codon, frame shifts, and exon length were also considered. Evolutionary conservation of many of these sequences was also included: although this feature is not one that mouse cells can perceive, it could aid a researcher in interpreting the code. On top of all this these features were considered separately for seven different regions: the alternatively spliced exon and 300 nt of its intronic flanks plus the upstream and downstream exons and their proximal intronic flanks. Note that these last four regions can be located thousands of nucleotides away from the exon in question. The separate consideration of these seven regions multiplies the number of features tracked. While tissue specific splicing motifs have been discovered by genomic analysis in the past (e.g.,9), this study stands out its comprehensiveness and in its consideration of distant locations.

About 200 of the original 1014 features proved to be useful. This filtered list includes confirmatory assignments for PTB and Nova binding sites, for example, but it also suggests unexpected roles for the density of many short sequences and, intriguingly, for sequences residing in the far-flung adjacent exon regions. Importantly, in a post-processing step, the authors could identify many pairs of features that significantly co-occurred, suggestive of specific molecular interactions. Overall, the results provide a list of players whose roles can now be followed up with mechanistic studies. The list also allows an exploration of the effect on splicing of SNPs that disrupt important features, a direction that could prove relevant to human disease. Even at this early stage, the authors were able to come up with evidence for increased gene expression in embryonic stem cells via the exclusion of alternatively spliced “killer” exons that reduce mRNA levels in adult tissue. Furthermore, the methodology itself can be applied to learn other codes and can be adopted and honed by others. While this comprehensive work has produced an important advance, there is more to be done. An improved code would provide quantitative predictions of exon inclusion rather than just directionality. And additional wet validation experiments to test the importance of features will be needed before the conclusions based on statistics can be accepted with confidence.

The strategy used by Barash et al. was not aimed at a general code for exon definition but rather at a code for alternative splicing, the difference in splicing behavior of a given exon in two different environments. Although there may be differences in how alternative exons are defined10, it would be surprising if many of the features identified here do not turn out to reflect basic mechanisms in splice site recognition. Indeed, the comparison of two different states (tissues) can help pinpoint such factors. Perhaps the most important message from this work is that each exon does not march to the beat of a different drummer, but gets spliced through a complex but knowable orchestration based on a large but limited set of instruments.

References

1. Barash Y, et al. Deciphering the splicing code. Nature. 2010; 465:53–59. [PubMed: 20445623]
2. Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. Rna. 2008; 14:802–813. [PubMed: 18369186]
3. Chasin LA. Searching for splicing motifs. Adv Exp Med Biol. 2007; 623:85–106. [PubMed: 18380342]
4. Fu XD. Towards a splicing code. Cell. 2004; 119:736–738. [PubMed: 15607969]
5. Trifonov EN. Interfering contexts of regulatory sequence elements. Comput Appl Biosci. 1996; 12:423–429. [PubMed: 8996791]
6. Munoz MJ, de la Mata M, Kornblihtt AR. The carboxy terminal domain of RNA polymerase II and alternative splicing. Trends Biochem Sci. 2010
7. Tilgner H, et al. Nucleosome positioning as a determinant of exon recognition. Nat Struct Mol Biol. 2009; 16:996–1001. [PubMed: 19684599]
8. Luco RF, et al. Regulation of alternative splicing by histone modifications. Science. 2010; 327:996–1000. [PubMed: 20133523]
9. Das D, et al. A correlation with exon expression approach to identify cisregulatory elements for tissue-specific alternative splicing. Nucleic Acids Res. 2007; 35:4845–4857. [PubMed: 17626050]
10. Xue Y, et al. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. Mol Cell. 2009; 36:996–1006. [PubMed: 20064465]

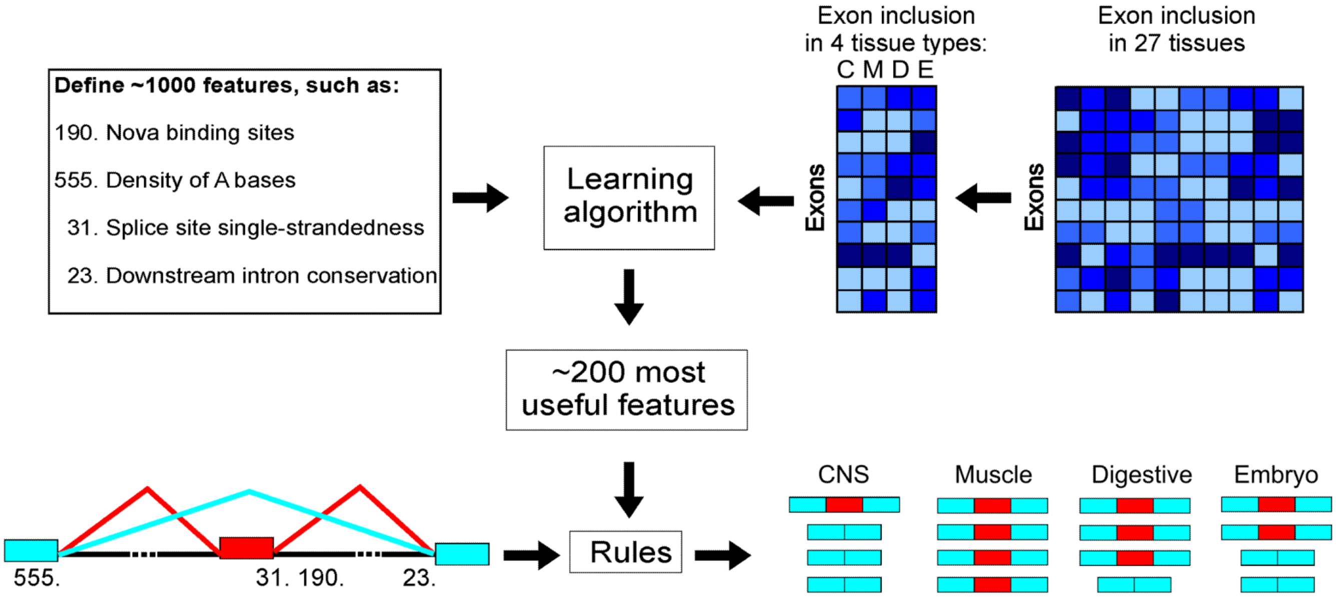
Define ~1000 features, such as:
190. Nova binding sites
555. Density of A bases
31. Splice site single-strandedness
23. Downstream intron conservation
Learning algorithm
Exon inclusion in 4 tissue types:
C M D E
Exons
Exon inclusion in 27 tissues
Exons
~200 most useful features
555.
31. 190.
23.
Rules
CNS
Muscle
Digestive
Embryo

Fig. 1.

Scheme for associating RNA sequence features with splicing outcomes. Top left: More than 1000 diverse features were used; the examples shown here were chosen to illustrate their diversity. Each feature was also defined by the region in which it occurs, as indicated on the map on the lower left, where the alternatively spliced exon is red. Upper right: Exon inclusion data were originally measured in 27 mouse tissues or cell lines using microarrays and then consolidated into four tissue types: C, central nervous system; M, striated and cardiac muscle; D, digestion related tissues; E, embryonic tissue and stem cells. A machine learning algorithm was devised to associate particular features with particular splicing outcomes; the latter being categorized as increased exon inclusion, increased exon exclusion, or no difference in comparing two tissue types. After training on a set of ~3000 exons, the algorithm was able to reliably predict these splicing outcomes in a set of test exons.