Splicing by cell type

Mauricio A Arias, Shengdong Ke & Lawrence A Chasin

A comprehensive study identifies sequence features that predict tissue-specific alternative splicing.

The rules governing exon splicing 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 analyzed to predict tissue-specific alternative splicing patterns. By using a multitude of features to describe an RNA molecule and focusing on cell-specific splicing decisions, the authors have provided a much richer picture of the code underlying alternative splicing than has been achieved previously.

In contrast to transcription and translation, in which the flow of information from DNA to pre-mRNA and from mRNA to protein is governed by simple codes, the processing of pre-mRNA to mRNA is less straightforward. Extracting exons from pre-mRNA and splicing them together to create mRNA requires, first and foremost, a mechanism for distinguishing exons and introns. Intron recognition always takes place during the splicing event itself, which is catalyzed by the large spliceosomal machinery comprising five RNA molecules and >100 proteins. In contrast, exon recognition is thought to occur before the splicing reaction. The main evidence for this is that disruption of an individual splice site most often leads to the entire exon being skipped.

How early exon recognition takes place is not well understood. The sequences immediately surrounding the splice sites themselves do not contain enough information to demarcate the borders of exons. Several lines of evidence have shown that additional information exists in short degenerate sequence motifs that lie both within and outside the exons. These genetic elements have been shown to interact with specific RNA-binding proteins to either enhance or silence splicing, but the underlying mechanisms have remained elusive. The composition, location and function of these sequence elements have been called the 'splicing code'[2–5].

Deciphering the splicing code is more complicated than analyzing the linear arrangement of these sequence elements, for several reasons. First, RNA can fold into intricate three-dimensional structures, driven mostly by base pairing between different regions of the molecule. The availability of a pre-mRNA sequence to bind an RNA-binding protein therefore depends on its structure. Pre-mRNA structure itself could also play a direct role in splicing. Second, as 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 of transcription to allow a splice site to be recognized[6]. Third, chromatin structure is emerging as a possible modulating factor in splicing (e.g., refs. 7 and 8). Thus, the splicing code can involve DNA sequences as well as RNA.

The situation is even more complicated because the splicing code can produce multiple outcomes in a given cell type and can be interpreted differently in different cellular environments. The result is alternative splicing, with the same gene giving rise to multiple mRNA isoforms and their corresponding protein isoforms. Although most exons are spliced constitutively—that is, included with near 100% efficiency in all mature mRNA molecules produced in all tissues examined—a large minority are alternatively spliced, such that almost all mammalian genes undergo some alternative splicing. Alternative splicing can generate a proteome that is much larger than the transcriptome, thereby explaining the relative complexity of higher organisms without much of a difference in genome size. Tissue-specific alternative splicing adds another layer to the splicing code, with differences between tissues presumably mediated by different repertoires or levels of splicing factors or chromatin structures. The code for tissue-specific alternative splicing may be part and parcel of the general code or distinct from it, or the two may overlap.

The study of Barash *et al.*[1] tackles the tissue-specific splicing code through a collaboration between computational and experimental researchers. The authors' strategy was to reveal the elements of the code by associating the presence of sequence 'features' with splicing outcomes (**Fig. 1**). The latter, determined by high-throughput microarray measurements of mRNA levels, comprised 3,665 alternatively spliced exons in 27 mouse cells and tissues. The complexity of the problem was then reduced in two ways. First, the 27 samples were grouped into four tissue categories (CNS, muscle, digestion and the embryo) for comparison. Second, relative percent inclusion levels were made discrete as three probabilities: increased, decreased or unchanged inclusion in a particular tissue compared to a baseline. 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 ~90% was achieved, attesting to the validity of the method.

The collection of sequence features is perhaps the heart of this study. The authors compiled a list of 1,014 diverse features using data in the literature and their own intuition. Most of the features were based on oligomeric sequences discovered in various types of experiments—for example, sets of predicted and validated hexamer sequences from statistical analysis of the transcriptome, ligand sequences for splicing factors and positional weight matrices for sequences derived by functional selection. But the feature list also included the density of all possible base trimers, dimers and even single bases. RNA structure was taken into account as predicted single-strandedness around regions such as the splice sites. Splice site scores, the creation of premature stop codons, frame shifts, exon length and evolutionary conservation were also included. In addition, the 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. 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. Whereas tissue-specific splicing motifs have been discovered by genomic analysis in the past (e.g., ref. 9), this study stands out for its comprehensiveness and its inclusion of distant locations.

About 200 of the original 1,014 features proved to be useful in predicting alternative

Mauricio A. Arias, Shengdong Ke and Lawrence A. Chasin are in the Department of Biological Sciences, Columbia University, New York, New York, USA.
e-mail: lac2@columbia.edu

splicing. This filtered list includes confirmatory assignments for binding sites of the polypyrimidine tract–binding protein and the Nova splicing factor, 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 single-nucleotide polymorphisms 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 through the exclusion of alternatively spliced 'killer' exons that reduce mRNA levels in adult tissue. Furthermore, the method itself can be applied to understand codes for processes other than splicing.

Although this comprehensive study represents an important advance, there is more to be done. An improved code would provide quantitative predictions of exon inclusion rather than just directionality. Additional wet validation experiments to test the importance of features would allow conclusions based on statistics to be accepted with confidence. The use of RNA-seq data to measure exon inclusion should improve the accuracy of the code. Finally, tissue-specific levels of RNA-binding proteins, RNA-binding-protein occupancy and nucleosome position and modification may provide additional useful information.

The strategy of Barash *et al.*[1] was not aimed at determining a general code for exon definition but rather a code for alternative splicing—the difference in the splicing of a given exon in two different environments. Although there may be differences in how alternative exons are defined[10], 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 is spliced through a complex but knowable system based on a large but definable set of features.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Barash, Y. *et al. Nature* **465**, 53–59 (2010).
2. Wang, Z. & Burge, C.B. *RNA* **14**, 802–813 (2008).
3. Chasin, L.A. *Adv. Exp. Med. Biol.* **623**, 85–106 (2007).

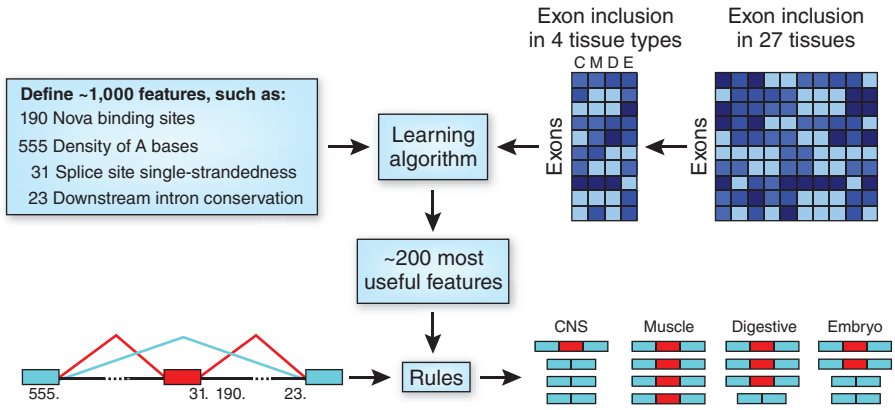

Figure 1 Scheme for associating RNA sequence features with splicing outcomes. Barash *et al.*[1] used >1,000 diverse sequence features (top left); the examples shown here were chosen to illustrate their diversity. Each feature was also defined by the region in which it occurs, as indicated on the map on the lower left, where the alternatively spliced exon is shown in red. 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 (upper right; darker shades represent higher exon inclusion levels). A machine learning algorithm was devised to associate particular features with particular splicing outcomes, the latter categorized as increased exon inclusion, increased exon exclusion or no difference between two tissue types. After training on a set of ~3,000 exons, the algorithm could reliably predict these splicing outcomes in a set of test exons.

4. Fu, X.D. *Cell* **119**, 736–738 (2004).
5. Trifonov, E.N. *Comput. Appl. Biosci.* **12**, 423–429 (1996).
6. Munoz, M.J., de la Mata, M. & Kornblihtt, A.R. *Trends Biochem. Sci.* (2010). doi:10.1016/j.tibs.2010.03.010.
7. Tilgner, H. *et al. Nat. Struct. Mol. Biol.* **16**, 996–1001 (2009).
8. Luco, R.F. *et al. Science* **327**, 996–1000 (2010).
9. Das, D. *et al. Nucleic Acids Res.* **35**, 4845–4857 (2007).
10. Xue, Y. *et al. Mol. Cell* **36**, 996–1006 (2009).

A synthetic DNA transplant

Mitsuhiro Itaya

The complete set of tools needed to synthesize a functional genome and transplant it into a mycoplasma cell opens up the possibility of mixing and matching natural and synthetic DNA to make genomes with new capabilities.

The recent creation of a new bacterium *Mycoplasma mycoides* JCVI-syn1.0 from an artificially constructed genome represents a technical *tour de force*. The accomplishment, described in a paper by Gibson *et al.*[1] of the J. Craig Venter Institute (JCVI; Rockville, MD, USA) published in *Science*, is the culmination of over a decade of effort to create a cell with an artificial genome. Although creation of a self-replicating cell using a computer as the starting point represents an important breakthrough for synthetic biology, several key details of the transplantation protocol remain to be established. Moreover, gaps in our knowledge of genome biology and the expense of producing whole genomes synthetically will likely limit wide adoption of the approach for the foreseeable future.

The synthetic biology group at JCVI has developed and released several basic methods[2–4] that together have made up incremental steps toward the ultimate aim of creating a synthetic genome that can then be transplanted into a recipient (so-called chassis) organism. In their present paper, Gibson *et al.*[1] now combine these methods and successfully apply them to design a particular mycoplasma strain that never existed before. The methods essentially comprise three major parts, as illustrated in

Mitsuhiro Itaya is at the Laboratory of Genome Design Biology, Institute for Advanced Biosciences, Keio University, Yamagata, Japan.
e-mail: mita2001@sfc.keio.ac.jp