NATURE|Vol 465|6 May 2010

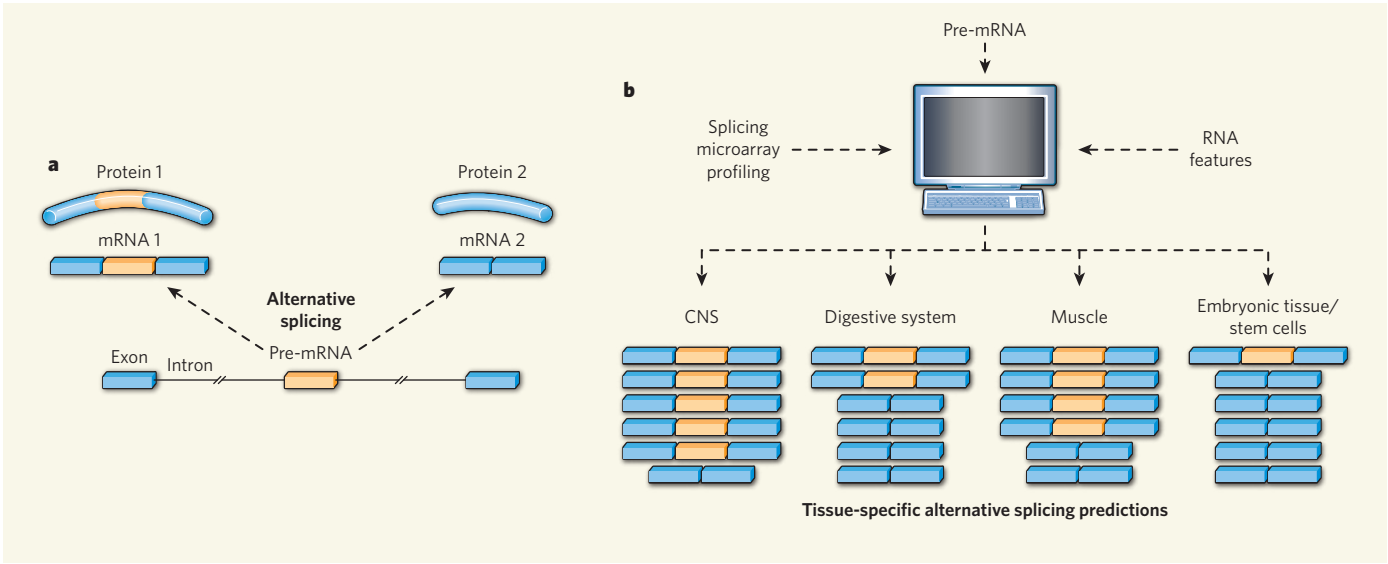

Figure 1 | Predicting pre-mRNA fate. a, Genomic DNA sequences are transcribed as messenger RNA precursors (pre-mRNA) containing exons and introns that can be processed by alternative pathways to generate different mRNAs encoding distinct proteins. **b**, Using data on alternative splicing obtained by microarray profiling of mRNAs from different tissues, as well as a compendium of regulatory sequences (RNA features), Barash *et al.*[1] identify combinations of features that can predict, for a given pre-mRNA, the ratio of alternatively spliced mRNAs in four different tissue types: central nervous system (CNS), digestive system, muscle and embryonic tissue/stem cells.

challenge, therefore, is to compute the algebra of a myriad of sequence motifs, and the mutual relationships between the regulatory factors that recognize them, to predict tissue-specific splicing.

To achieve this, Barash *et al.*[1] provided a computer with two types of information (Fig. 1b). The researchers gathered microarray data evaluating the ratio between inclusion and skipping of more than 3,000 alternatively spliced exons in four types of mouse tissue. They also took advantage of the collective knowledge generated by the splicing research community to compile thousands of RNA sequence features corresponding to known binding sites for regulatory factors, as well as sequence motifs enriched around alternatively spliced RNA regions, even if their cognate regulatory factors remain unknown. Moreover, they considered characteristics of the exon/intron organization, their evolutionary conservation, the folded structure of the RNA chain and the relationships among all these elements. The computer was then asked to identify the combination of features that could best explain the experimentally determined tissue-specific selection of exons.

Considering the complexity of the system, the approach achieved notable successes. It correctly identified alternative exons, and predicted their differential regulation between pairs of tissue types with considerable accuracy. The code identified features whose distribution and frequent co-occurrence with other sequence elements is associated with tissue-specific regulation. This allows reinterpretation of the function of previously defined regulatory motifs and suggests previously unknown properties of known regulators as well as unexpected functional links between them. For instance, the code inferred that the inclusion of exons that lead to truncated proteins is a common mechanism of gene-expression control during the transition between embryonic and adult tissues.

Despite these successes, however, Barash and colleagues' work[1] may be better seen as revealing the first piece of a much larger Rosetta Stone required to interpret the alternative messages of our genomes. The expected wave of massive data sets generated by high-throughput technologies[6] should soon provide further inputs for improving the code. These include identification *in vivo* of binding sites for regulatory proteins by techniques such as cross-linking/immunoprecipitation (CLIP), extensive description of mRNA variants by high-throughput sequencing, and functional characterization of regulators by RNA interference screens.

The code is likely to work in a cell-autonomous manner and, consequently, may need to account for more than 200 cell types in mammals. It will also have to deal with the extensive diversity of alternative-splicing patterns beyond simple decisions of single exon inclusion or skipping. The limited evolutionary conservation of alternative-splicing regulation (estimated to be around 20% between humans and mice) opens up the question of species-specific codes. Moreover, coupling between RNA processing and gene transcription influences alternative splicing, and recent data[7,8] implicate the packing of DNA with histone proteins and histone covalent modifications — the epigenetic code — in the regulation of splicing. The interplay between the histone and the splicing codes will therefore need to be accurately formulated in future approaches. The same applies to the still poorly understood influence of complex RNA structures on alternative splicing.

Deciphering the genetic code allowed the identification of protein-coding genes and thus provided a key conceptual framework for understanding genome organization. An important measurement of the value of Barash and co-workers' paper[1] will be its usefulness in interpreting the output of genes in genome-sequencing projects and in rationalizing changes in alternative splicing caused by natural sequence variation or underlying pathological conditions. Another key assessor of this[1] and other codes of post-transcriptional regulation[9] will be their amenability to converting large data sets of intriguing relationships between sequence motifs into testable hypotheses that will help to unravel the underlying mechanisms, the code's molecular fabric. ■

J. Ramón Tejedor and Juan Valcárcel are at the Centre de Regulació Genòmica and Universitat Pompeu Fabra, Barcelona, Spain. Juan Valcárcel is also at the Institució Catalana de Recerca i Estudis Avançats, Dr. Aiguader 88, 08003 Barcelona, Spain.
e-mail: juan.valcarcel@crg.es

1. Barash, Y. *et al. Nature* **465,** 53–59 (2010).
2. Nirenberg, M. *Trends Biochem. Sci.* **29,** 46–54 (2004).
3. Nilsen, T. W. & Graveley, B. R. *Nature* **463,** 457–463 (2010).
4. Chasin, L. A. *Adv. Exp. Med. Biol.* **623,** 85–106 (2007).
5. Chen, M. & Manley, J. L. *Nature Rev. Mol. Cell Biol.* **10,** 741–754 (2009).
6. Blencowe, B. J., Ahmad, S. & Lee, L. J. *Genes Dev.* **23,** 1379–1386 (2009).
7. Luco, R. F. *et al. Science* **327,** 996–1000 (2010).
8. Fox-Walsh, K. & Fu, X.-D. *Dev. Cell* **18,** 336–338 (2010).
9. Piqué, M., López, J. M., Foissac, S., Guigó, R. & Méndez, R. *Cell* **132,** 434–448 (2008).

Addendum
Gökhan S. Hotamisligil, a co-author of the News & Views article 'Metabolism: Host and microbes in a pickle' (*Nature* **464,** 1287–1288; 2010) declared a competing financial interest that was not noted in the print or PDF versions of the article. The declaration can be found in the full-text online version.