Piecing together the significance of splicing

Rotem Sorek and Mor Amitai

Alternative splicing increases protein diversity by allowing multiple, sometimes functionally distinct, proteins to be encoded by the same gene. It can be specific to tissues, stress conditions, and developmental and pathological states. In many cases, it serves as an on/off regulation mechanism by introducing a premature stop codon[1]. We still do not understand how alternative splicing is regulated, but the following fact is now quite clear: in metazoans, it happens very often.

Computational analysis of expressed sequences can teach us a great deal about alternative splicing, because aligning expressed sequences of different splice variants of the same gene usually results in a typical gapped pattern of alignment. The recent information explosion in nucleotide databases gives us the possibility of analysis at the transcriptome (the set of messenger RNAs) level. We can take the entire human expressed sequence tags (ESTs) database (http://www.ncbi.nlm.nih.gov/dbEST/) (currently containing more than 3.1 million ESTs), together with the known complementary DNAs and genomic sequence data, and cluster them by alignment overlaps.

Such large-scale analyses were conducted in the past few years by four independent groups using different methods[2–5]. They estimated that 33–59% of human genes have at least two splice variants, with the highest estimation being the most recent one[2]. All four groups also pointed out that these figures are probably an underestimation, because the EST database does not cover the entire repertoire of tissues or developmental states, and precautions taken to avoid false positives were extremely stringent. Because the latest estimation of the total number of human genes is 30,000–40,000 (ref. 2), one must bear in mind that at least 10,000 of our genes, and probably many more, undergo alternative splicing.

Understanding alternative splicing and gaining knowledge of the transcriptome are crucial for the design and interpretation of expression profiling experiments, in particular DNA chip experiments. Such experiments enable comparisons between transcriptomes of different cell types or under different conditions. Designing DNA chips that will effectively report on the transcriptional levels of genes must take into account their alternative splicing patterns, even if alternative splicing is not the subject being studied.

To demonstrate this, let us assume a putative gene *X* that has three exons (A, B, and C) and two splice variants, ABC and AC. To design a chip that will measure the transcriptional levels of gene *X* in different tissues, we must use a probe from exons A or C to which both variants will hybridize in the assay. Taking the probe from exon B will cause the hybridization of variant ABC only, and will not correctly measure the transcriptional level of the gene. The accuracy of the experiment can be increased by measuring the level of each variant separately; hence, two probes will have to be taken, one from exon B to measure variant ABC, and one from exon A or C to measure both variants.

Another illustration of the importance of awareness of alternative splicing comes from the field of gene prediction. One thing gene prediction programs do not predict is alternative splicing, because sequences that regulate alternative splicing are generally unknown. Because alternative splice sites often correspond weakly to the splice consensus sites, gene prediction programs will probably frequently fail to identify alternative exons or introns.

Alternatively spliced genes are likely to take center stage as drug targets, therapeutic agents, and diagnostics markers in the next decade. First, there are many splice variants of pharmaceutically important genes that have been detected but not yet studied in depth. The function of the known variant gives us a clue to the function of the new variant, especially if we know which domain was added or removed. For example, we have identified some 60 kinase enzymes that undergo alternative splicing that eliminates their catalytic domains (E. Levanon *et al.*, unpublished data). Although many of them have not been biochemically studied, our educated guess is that they function as competitive inhibitors of the known kinases.

Second, it has been estimated that 15% of the point mutations that cause genetic diseases in humans alter the normal splicing pattern[6]. Splice variants that are disease specific can be excellent diagnostic markers for these and other human diseases, being easily identifiable by PCR reactions.

Alternative splicing is no longer considered an esoteric twist of nature. Articles with the phrase "alternative splicing" or "splice variants" in their title or abstract are published at the rate of two a day (according to *Medline* query). In many ways, the concept is breaking our iron-clad rules: exons are not always exons, and introns are sometimes expressed.

Indeed, the very definition of "gene" should be reconsidered in light of discoveries of unusual alternative splicing events, such as the one yielding a novel splice variant of PSA (prostate specific antigen)—the standard prostate cancer marker[7]. This variant shares with PSA only the first exon, which encodes only a signal peptide, leaving the two mature proteins with no common protein sequence. The only connection between them is that they are coded by the same genomic region and probably share the same transcriptional regulation (A. David *et al.*, unpublished data).

Even more extreme is the example of the p19 and p16 protein products of the *INK4a/ARF* locus[8]. The two transcripts are synthesized from different promoters and have different first exons, but share exons 2 and 3, and are encoded in two distinct reading frames, in a process that yields two entirely different protein products. Although the p19 and p16 proteins are clearly the products of the same genomic locus, can we say that these two unusual and entirely distinct splice variants are coded by the same gene?

Clearly, increasing protein diversity does not simply correlate with increasing gene number. It is dependent both on the number of genes in the genome and on the rate of alternative splicing of those genes. Work is now needed to characterize in greater detail the molecular basis for this process and its regulation. This will likely uncover a host of new targets for drug discovery, yield new diagnostic markers for disease, and perhaps even help us unravel the mechanisms underlying biological complexity.

Rotem Sorek is a scientist (rotem.sorek@cgen.com) and Mor Amitai (mor.amitai@cgen.com) is the president and chief executive officer at Compugen, 72 Pinchas Rosen St., Tel-Aviv, Israel, 69512.

1. Smith, C.W. & Valcarcel, J. *Trends Biochem. Sci.* **25**, 381–388 (2000).
2. International Human Genome Sequencing Consortium. *Nature* **409**, 860–921 (2001).
3. Brett, D. *et al. FEBS Lett.* **474**, 83–86 (2000).
4. Mironov, A.A., Fickett, J.W. & Gelfand, M.S. *Genome Res*. **9**, 1288–1293 (1999).
5. Croft, L. *et al. Nat. Genet.* **24**, 340–341 (2000).
6. Cooper, T.A. & Mattox, W. *Am. J. Hum. Genet.* **61**, 259–266 (1997).
7. Diamandis, E.P. *Trends Endocrinol. Metab.* **9**, 310-316 (1998).
8. Sharpless, N.E. & DePinho, R.A. *Curr. Opin. Genet. Dev*. **9**, 22-30 (1999).