Cell, Vol. 103, 367–370, October 27, 2000, Copyright ©2000 by Cell Press

Protein Diversity from Alternative Splicing: A Challenge for Bioinformatics and Post-Genome Biology

Minireview

Douglas L. Black*
Howard Hughes Medical Institute
University of California, Los Angeles
MRL 5-748
675 Charles E. Young Dr. South
Los Angeles, California 90095

The human genome sequence can be thought of as a picture of the human organism. However, like an impressionist painting, the genome is a very large canvas whose details become fuzzy when you look closely. A fully detailed image of a complex organism requires knowledge of all of the proteins and RNAs produced from its genome. This is the impetus for proteomics, the study of the complete protein sets of organisms. Due to the production of multiple mRNAs through alternative RNA processing pathways, human proteins often come in multiple variant forms. Because of our ignorance of the rules governing splice site choice, today's tools for analyzing genomic sequence provide a picture of these gene products that is highly indistinct. Our ability to define the product RNA and protein structures encoded within genomic sequence will need to improve greatly before a complete genome sequence can tell us the fine details of an organism's protein constitution.

Splicing Analysis as a Fine Brush on the Genome

Alternative splicing is seen in nearly all metazoan organisms as a means for producing functionally diverse polypeptides from a single gene (Lopez, 1998). It is especially common in vertebrates; alignment of EST sequences and mapping the resulting mRNA families to the human genome provided a minimum estimate that 35% of human genes show variably spliced products (Croft et al., 2000 and references therein). However, since these ESTs derive from a limited number of tissues or developmental states, and cover only a limited portion of each mRNA, the true percentage is likely much higher. Moreover, it is common to see genes with a dozen or more different transcripts. There are also remarkable examples of hundreds and even thousands of functionally divergent mRNAs and proteins being produced from a single gene. In the human genome, such protein-rich genes include the Neurexins, n-Cadherins, calcium-activated potassium channels, and others. Thus, the estimated 35,000–80,000 genes in the human genome could easily produce several hundred thousand different proteins, and possibly more.

Variation in mRNA structure takes many different forms (Lopez, 1998; Smith and Valcarcel, 2000). Exons can be spliced into the mRNA or skipped. Introns that are normally excised can be retained in the mRNA. The positions of either 5′ or 3′ splice sites can shift to make exons longer or shorter. In addition to these changes in splicing, alterations in transcriptional start site or polyadenylation site also allow production of multiple mRNAs from a single gene. All of these changes in mRNA structure can be regulated in diverse ways, depending on sexual genotype, cellular differentiation, or the activation of particular cell signaling pathways.

The effect of altered mRNA splicing on the structure of the encoded protein is similarly diverse (Lopez, 1998; Smith and Valcarcel, 2000). In some transcripts, whole functional domains can be added or subtracted from the protein coding sequence. In other systems, the introduction of an early stop codon can result in a truncated protein or an unstable mRNA (Morrison et al., 1997). Alternative splicing is also commonly used to control the inclusion of particular short peptide sequences within a longer protein. These optional peptide sequence cassettes range from one to hundreds of amino acids in length, and have very specific effects on the activity of a protein product. Changes in splicing have been shown to determine the ligand binding of growth factor receptors and cell adhesion molecules, and to alter the activation domains of transcription factors (Lopez, 1998; Smith and Valcarcel, 2000). In other systems, the splicing pattern of an mRNA determines the subcellular localization of the encoded protein, the phosphorylation of the protein by kinases, or the binding of an enzyme by its allosteric effector. Determining how these sometimes subtle changes in sequence affect protein function is a crucial question in many different problems in developmental and cell biology including control of apoptosis (Jiang and Wu, 1999), tumor progression (Herrlich et al., 1993), neuronal connectivity (Schmucker et al., 2000), and the tuning of cell excitation and cell contraction.

A recent discovery in *Drosophila* is both a fascinating example of the subtle structural changes that can be made in a protein and a remarkable demonstration of the number of proteins that can be produced from a single gene using alternative splicing. *Drosophila* DSCAM protein was cloned as an axon guidance receptor responsible for directing growth cones to their proper target in Bolwig's nerve of the fly (Schmucker et al., 2000). The isolated DSCAM cDNAs showed several positions of heterogeneity attributable to alternative splicing. However, comparison of these cDNAs to the DSCAM genomic sequence held a surprise. In addition to the exons encoding the isolated cDNAs, dozens of different homologous exons were also present in what would be the DSCAM primary transcript. Exons 4, 6, 9, and 17 are each encoded by an array of potential alternative exons (Figure 1). These exons are used in a mutually exclusive manner, where there are 12 alternatives for exon 4, 48 alternatives for exon 6, 33 alternatives for exon 9, and 2 alternatives for exon 17. If all combinations of these exons were used, the single *DSCAM* gene would produce 38,016 different DSCAM proteins! Cloning and sequencing 50 different random cDNAs identified 49 different combinations of exons 4, 6, and 9. Thus, even if not all exon combinations are allowed, it is clear that this one gene produces many thousands of protein products.

Although it is not yet known how they affect function, the changes in DSCAM protein structure brought about by these changes in splicing are interesting. Exons 4

* E-mail: dougb@microbio.lifesci.ucla.edu

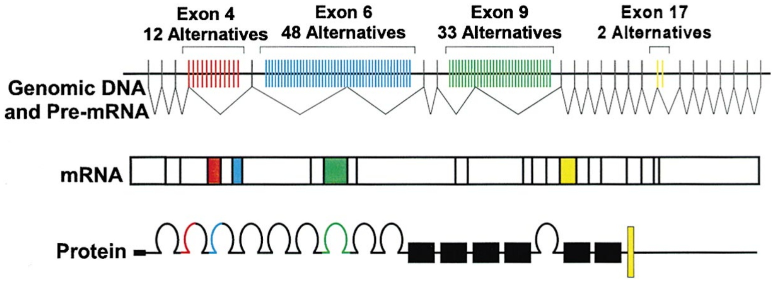

Figure 1. 38,016 Shades of DSCAM

The *DSCAM* gene (top) is 61.2 kb long and after transcription and splicing produces a 7.8 kb, 24 exon mRNA (middle). Exons 4, 6, 9, and 17 are encoded as arrays of mutually exclusive alternative exons. Each mRNA will contain one of 12 possible alternatives for exon 4 (in red), one of 48 for exon 6 (blue), one of 33 for exon 9 (green) and one of 2 for exon 17 (yellow). In the final protein product (bottom), exon 4 encodes the amino-terminal half of Ig domain 2. Exon 6 encodes the same portion of Ig domain 3, exon 9 encodes all of Ig domain 7, and 17 encodes the transmembrane domain. If all possible combinations of single exons 4, 6, 9, and 17 are used, the *DSCAM* gene produces 38,016 different mRNAs and proteins.

and 6 encode the N-terminal half of the 2nd and 3rd immunoglobulin domains within the extracellular portion of the receptor. The multiple forms of exon 9 each encode an entire Ig domain 7. Each exon 17 encodes an alternate transmembrane domain. A similar splicing change in an Ig domain of the FGF Receptor 2 mRNA determines which of several growth factors the receptor will respond to. Thus, although similar in overall protein structure, the different DSCAM receptors are likely to have important differences in activity. The implications of this diversity in axon guidance receptors for neuronal development have been discussed elsewhere (Schmucker et al., 2000), but the gene itself provides a spectacular indication of how much protein diversity can be generated by alternative splicing. The *Drosophila* genome contains approximately 13,600 identified genes (Adams et al., 2000), whereas this single gene can produce nearly three times that number of proteins. It has been a puzzle that an organism as complex as a fly would need so few genes to describe all of its functions. It seems clear that due to alternative splicing the gene number is not an estimate of the protein complexity of the organism.

However fascinating this diversity of products, we are at a loss to explain what might control DSCAM splicing, for the mechanisms of alternative splicing regulation are poorly understood. During the course of the splicing reaction, each intron in a pre-mRNA is assembled into a spliceosome complex, where the splice sites at the intron ends are brought together and the cleavage and ligation reactions are catalyzed (Staley and Guthrie, 1998). The spliceosome is a large particle made of a set of 5 small nuclear ribonucleoproteins, the U1, U2, U4, U5, and U6 snRNPs, as well as a number of important auxiliary proteins. The splice sites at the intron/exon junctions adhere to particular consensus sequences. Early in spliceosome assembly, the 5′ splice site at the 5′ end of the intron is bound by the U1 snRNP. One of the non-snRNP proteins, the U2 Auxiliary Factor (U2AF), binds to the polypyrimidine tract of the 3′ splice site and stimulates U2 snRNP binding upstream (Figure 2). At this point in the assembly of the spliceosome, the splice sites have apparently been chosen and paired to define the excised intron. Later steps in the assembly bring in the other snRNPs and rearrange their structure to put the splice sites into the active center of the spliceosome and allow for catalysis. The alteration of splice site choice is an alteration in the assembly of the early spliceosome complex that changes the end points of an excised intron and hence the structure of the final mRNA.

The choice of splice sites is thought to be directed by proteins that bind to special non–splice site RNA elements and then enhance or repress spliceosome assembly nearby (Figure 2) (Lopez, 1998; Smith and Valcarcel, 2000). Splicing regulators include members of the SR family of proteins as well as factors generally classed as hnRNP proteins. The best understood are the SR proteins which, among other activities, can bind to exonic splicing enhancer elements and stimulate spliceosome assembly at adjacent splice sites, possibly through direct interactions with U2AF and the U1 snRNP. Along a given pre-mRNA, a great many different SR and hnRNP proteins will bind to many short sequence elements producing a complex RNP structure (Krecic and Swanson, 1999; Lopez, 1998; Smith and Valcarcel, 2000). Most systems of alternative splicing appear to be highly combinatorial, with multiple positive and negative factors and elements influencing the final level of a spliced transcript. This complexity makes the molecular characterization of splicing regulation challenging. So far, only a few systems of regulated splice site choice have been genetically or biochemically dissected, and most regulatory proteins and sequence elements are not yet identified.

The general lack of understanding of how cells choose splice sites, and how they change particular choices, makes it difficult to identify exons and predict splicing patterns within genomic sequence. Functional splice sites do not always match the consensus sequences well. Conversely, there are many cryptic sites in the genome that match the consensus but are not normally recognized by the splicing apparatus. The sequence surrounding a splice site, as well as its match to the consensus, strongly affects its recognition. Introns can be very long, whereas exons are generally short, and one feature affecting the recognition of a splice site is the presence of the opposite site across the exon. Thus, splicing at a 3′ splice site in one intron can be stimulated by the 5′ splice site across the exon in the downstream intron. This has given rise to the concept of exon definition, where splice sites are thought to be recognized as one end of an exon unit, rather than as individual sites separated by a long intron (Berget, 1995). However, even by looking for splice sites in exon pairs, it is often difficult to distinguish real from cryptic sites; it is clear that sequences outside the splice sites themselves strongly affect their use. Current gene finding programs identify exons based on multiple properties such as coding capacity combined with matches to the consensus splice sites (Burge and Karlin, 1997; Haussler, 1998 and refer-

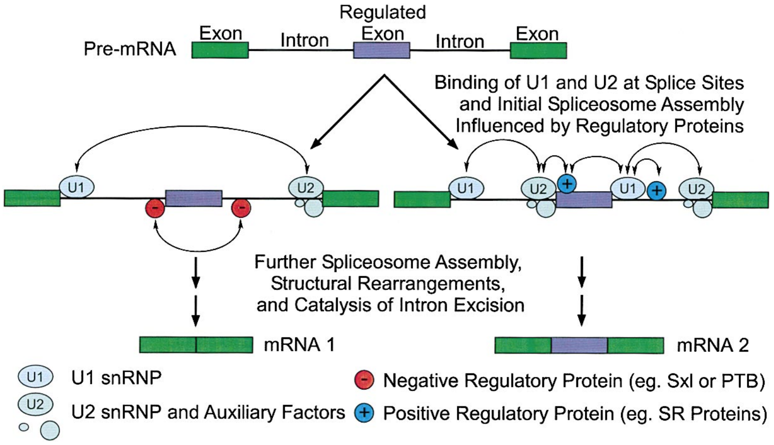
Pre-mRNA
Exon
Intron
Regulated Exon
Intron
Exon
Binding of U1 and U2 at Splice Sites and Initial Spliceosome Assembly Influenced by Regulatory Proteins
U1
U2
U1
U2
+
U1
+
U2
-
-
Further Spliceosome Assembly, Structural Rearrangements, and Catalysis of Intron Excision
mRNA 1
mRNA 2
U1 U1 snRNP
U2 U2 snRNP and Auxiliary Factors
- Negative Regulatory Protein (eg. Sxl or PTB)
+ Positive Regulatory Protein (eg. SR Proteins)

Figure 2. Spliceosome Assembly on Regulated and Unregulated Splice Sites

A pre-mRNA (top) is spliced according to where the spliceosome assembles and defines its introns and exons. The U1 snRNP binds to 5′ splice sites. The U2 snRNP and auxiliary proteins, including U2AF and SF1, bind to the branchpoint and the 3′ splice site. The binding of these initial components and the assembly of the early spliceosome complexes is thought to define the intron to be excised. In later steps, the full spliceosome forms and catalyzes the cleavage and ligation reactions at the earlier defined splice sites. The spliceosome can assemble between exon 1 and exon 3 to excise a single large intron and form mRNA 1. Alternatively, two spliceosomes can excise two smaller introns, thus including a new exon in the mRNA (mRNA 2). These outcomes can be determined by negative regulatory proteins (red) that either prevent U1 or U2 binding at particular sites (as shown), or block spliceosome assembly after U1 or U2 binding. Regulatory proteins can also act positively (blue) to enhance spliceosome assembly at sites that are otherwise recognized poorly. Most systems of alternative splicing seem to be controlled by multiple regulatory proteins that may exert both positive and negative control. Splicing patterns can also be affected by other factors, including RNA secondary structure and transcription rate.

ences therein). Exons containing clear open reading frames are most easily recognized, but at the ends of these exons the splice sites can be difficult to predict when there are multiple consensus sequence matches in the region. Exon prediction becomes even less successful with short exons. Microexons, encoding from one to a dozen or so amino acids, are common, but are difficult to identify in genomic DNA because of their lack of coding capacity. Until we know more about how cells recognize splice sites, it will be difficult to write software to predict the exonic structure of genes. This is a major limitation in our ability to annotate the genome.

The Long Bioinformatics View of Splicing

With a painting, although the close view may be indistinct, the view of the whole can resolve itself into a discernable picture. Similarly, the global view of the genome will provide a unique vantage for understanding splicing. Since only a few systems of alternative splicing are likely to be analyzed in biochemical detail at least in the near term, bioinformatics approaches will be important in predicting alternative splicing patterns from genomic sequence. Simple sequence comparisons are a first step. Microexons and intronic splicing regulatory regions are often much more highly conserved between species than other intron sequence (see for example Thackeray and Ganetzky, 1995). Comparison of the mouse and human or the *D. melanogaster* and *D. virilis* genomes will yield a great deal of information on exon location and on splicing regulatory sequences.

Work is also underway to align the EST and cDNA databases with genome sequences (Kent and Zahler, 2000; Wolfsberg and Landsman, 1997). Although this alignment approach again has difficulties in identifying short exons and exon termini correctly (Florea et al., 1998), it can in principle identify all the splicing occurring within the sequenced portion of existing cDNAs. At the moment, this is only a small portion of the splicing events in the genome, but projects to generate large databases of full-length cDNA sequences will greatly improve its coverage (Strausberg et al., 1999; Rubin et al., 2000). These efforts are unlikely to identify many low abundance mRNAs and will miss potentially important products. However, they will provide an initial reference splicing pattern for many genes. Such a large-scale identification of constitutive and regulated exons should also give us a great deal of information on the features and sequences that distinguish actual exons from the cryptic splice sites not normally recognized by the splicing apparatus. In addition to improving algorithms for predicting spliced segments within genomic sequence, this information will give insight into the mechanisms of alternative splicing and help experimentalists make sense of their complex biochemical systems.

Much more difficult than identifying exons correctly will be predicting splicing regulatory patterns from genomic sequence. This issue can be seen in the *DSCAM* gene sequence where the alternative exons were relatively easy to identify through protein homology (Schmucker et al., 2000). However, the same sequence gives no clues about how these exons are regulated or which exons might be used in particular cells. Alternative exons have binding sites for multiple regulatory proteins that often show only subtle variation between tissues, and the sequence elements that control exons of very different tissue specificity can look similar (Smith and Valcarcel, 2000). Thus, even where we identify important regulatory sequences and proteins for an exon, it will be difficult to predict the precise tissues or conditions that lead to splicing activation or repression.

The combinatorial nature of splicing regulation is similar to the control of transcription through promoter and enhancer elements and poses similar problems (Smith and Valcarcel, 2000). Many of the whole genome experimental approaches already being applied to transcriptional regulation can be informative for splicing as well (Young, 2000). Microarray technologies that allow the simultaneous assessment of the splicing of many exons within an RNA sample should prove particularly helpful. Unlike the detection of RNAs to measure whole transcript levels, these alternative splicing detector arrays will monitor the inclusion of particular exons in different populations of mRNAs. In one strategy, a position on the array would contain an oligonucleotide complementary to either a differentially included exon or to the exon/exon junction generated when this exon is skipped. The relative hybridization of a sample to these two sequences will give a measure of the relative inclusion of the exon. One limitation of this approach is that it does not give information correlating the exons within a single mRNA. For example, in the *DSCAM* gene, one would identify the alternatives for exons 4 and 6 that were present in a total mRNA sample, but which particular exons 4 and 6 were used together in the same mRNA molecule would not be discernable. Most importantly, however, such a technology permits one to examine the coordinate regulation of large groups of exons depending on development, cell type, or extracellular stimulus. This system-wide data about exon use may lead to the identification of sequence features that determine particular patterns of expression.

The presence or absence of particular exons in an mRNA can also be correlated with data on the expression patterns of potential splicing regulators, such as SR proteins or hnRNPs. Combined with conditional knockouts of these regulators, we can ask precise questions about what combinations of regulatory proteins are needed for particular exons. Although it is recognized as a significant problem in the understanding transcriptional regulation, the question of combinatorial control of splicing is only beginning to be addressed. Nevertheless, this understanding of how exons are regulated on a system-wide scale will ultimately be essential in interpreting genome sequence and predicting how and when certain proteins are produced from it.

Only a global view of splicing regulation combined with a detailed understanding of its mechanisms will allow us to paint a picture of an organism's total complement of proteins and of how this complement changes with development and the environment. Working toward this goal should keep bioinformatics researchers and molecular biologists busy for some time into the post-genome era.

Selected Reading

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). Science *287*, 2185–2195.

Berget, S.M. (1995). J. Biol. Chem. *270*, 2411–2414.

Burge, C., and Karlin, S. (1997). J. Mol. Biol. *268*, 78–94.

Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P., and Mattick, J.S. (2000). Nat. Genet. *24*, 340–341.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. (1998). Genome Res. *8*, 967–974.

Haussler, D. (1998). Trends Biochem. Sci., Supplementary Guide to Bioinformatics, pp. 12–15.

Herrlich, P., Zoller, M., Pals, S.T., and Ponta, H. (1993). Immunol. Today *14*, 395–399.

Jiang, Z.H., and Wu, J.Y. (1999). Proc. Soc. Exp. Biol. Med. *220*, 64–72.

Kent, W.J., and Zahler, A.M. (2000). Nucleic Acids Res. *28*, 91–93.

Krecic, A.M., and Swanson, M.S. (1999). Curr. Opin. Cell Biol. *11*, 363–371.

Lopez, A.J. (1998). Annu. Rev. Genet. *32*, 279–305.

Morrison, M., Harris, K.S., and Roth, M.B. (1997). Proc. Natl. Acad. Sci. USA *94*, 9782–9785.

Rubin, G.M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M., and Harvey, D.A. (2000). Science *287*, 2222–2224.

Schmucker, D., Clemens, J., Shu, J., Worby, C., Xiao, J., Muda, M., Dixon, J., and Zipursky, L. (2000). Cell *101*, 671–684.

Smith, C.W., and Valcarcel, J. (2000). Trends Biochem. Sci. *25*, 349–404.

Staley, J.P., and Guthrie, C. (1998). Cell *92*, 315–326.

Strausberg, R.L., Feingold, E.A., Klausner, R.D., and Collins, F.S. (1999). Science *286*, 455–457.

Thackeray, J.R., and Ganetzky, B. (1995). Genetics *141*, 203–214.

Wolfsberg, T.G., and Landsman, D. (1997). Nucleic Acids Res. *25*, 1626–1632.

Young, R. (2000). Cell *102*, 9–15.