# The importance of alternative splicing in the drug discovery process

Erez Y. Levanon and Rotem Sorek

**The publication of the sequence of the human genome revealed that the gene count in humans is much lower than previously estimated. Although textbooks usually place the number at 100,000, it is currently estimated that the human genome contains no more than 30,000 protein-coding genes. How can the great complexity of human life be explained by this number, which is less than twice the number of genes in the primitive worm *C. elegans*? The answer probably lies in the recent discovery that about half of all human genes undergo alternative splicing. This paper reviews the broad implications of alternative splicing for the drug-discovery process.**

**Erez Y. Levanon**
**Rotem Sorek***
Compugen Ltd.
72 Pinchas Rosen Street
Tel Aviv 69512, Israel.
*e-mail: rotem@compugen.co.il

▼ Mammalian genes are organized on the DNA in a typical exon–intron structure, with an average of 8.7 exons per gene [1]. Following the transcription of the DNA into pre-mRNA, the introns are cut out in a process called splicing. A huge RNA–protein complex, called a spliceosome [2,3], recognizes conserved sequences (splice sites) at the intron–exon boundaries and performs the actual splicing.

Alternative splicing takes place when the introns of a certain pre-mRNA can be spliced in more than one way, yielding several possible mature mRNAs from the gene (Fig. 1). Thus, one gene can produce several, sometimes functionally distinct, proteins (recently reviewed in [4,5]). The combinatorial nature of alternative splicing enables the genome to produce numerous transcripts out of a relatively small number of genes. One of the most extreme examples is the *Drosophila* Down syndrome cell adhesion molecule gene (*Dscam*), which can produce up to 38,016 different transcripts from the same pre-mRNA [6]. As about half of all mammalian genes are estimated to have more than one splice form [7–12], alternative splicing is a major factor accounting for the discrepancy between the size of the mammalian genome and the size of the proteome.

Alternative splicing is strictly regulated. Patterns of alternative splicing can be specific to tissue, developmental stage, condition or pathological state [5]. Different splice variants of the same gene can also appear in the same tissue and act as regulators of each other. For instance, a growing number of studies show that splice variants that lack an important domain of the protein can function as dominant negative proteins and competitively inhibit the activity of the full-length form. Thus, several apoptosis-related proteins, such as Bcl-x, Caspase-9, Ced-4 and Ich-1, have dominant negative splice variants that regulate their activity [13].

Aberrant splicing can result in pathological states. Approximately 15% of the single base-pair mutations that cause human genetic diseases are thought to be linked to pre-mRNA splicing defects [14]. The human mutations database [15] currently contains >3000 entries describing such mutations. In addition, some pathological states, such as cancer, cause aberrant splicing [16]. Thus, splice variants could serve as diagnostic markers for disease conditions.

It is becoming increasingly clear that successful drug discovery requires an understanding of complex biological contexts. Alternative splicing adds another important dimension of complexity to the proteomic world. Although information on alternative splicing has been accumulating at a rapid rate during the last four years [7–12], the core drug discovery processes entail techniques that cannot distinguish between splice variants and are therefore trapped in the old dogma of 'one gene, one mRNA, one protein'.

When the phenomenon of alternative splicing is overlooked, drug discovery processes are exposed to only a fraction of the actual proteomic world and therefore miss many potential protein targets. Even in experiments focused on a specific gene, results will be biased

1477-3627/03/$ – see front matter  PII: S1477-3627(03)02322-5

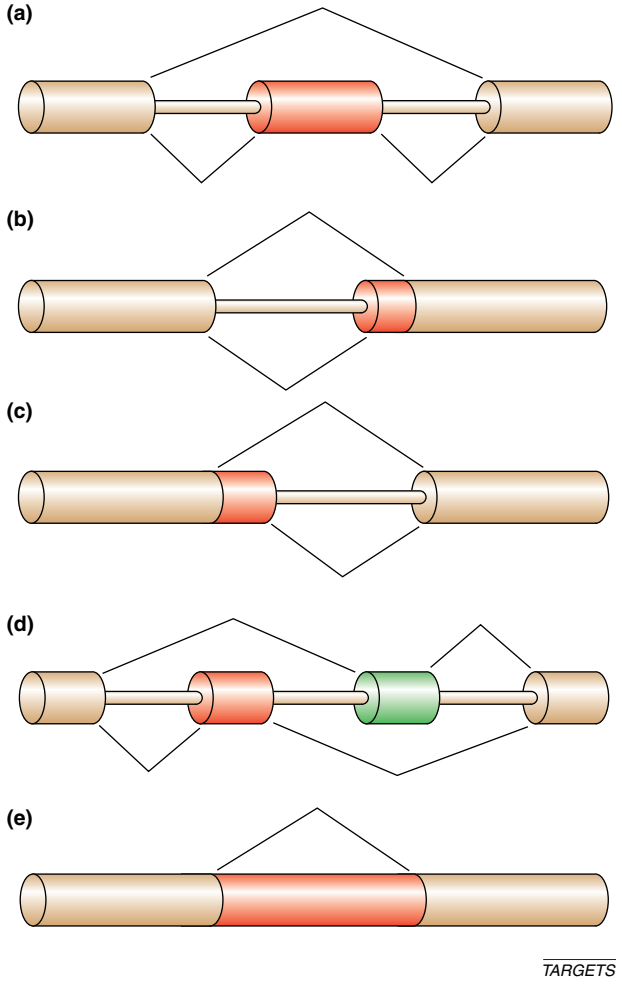


**Figure 1.** Possible patterns of alternative splicing. Exons are shown as cylinders; introns are thin lines. Triangles mark excision of introns. Beige exons are constitutively spliced; red and green exons are alternative. **(a)** Exon-skipping, also designated cassette-exon. **(b)** Alternative 3′ splice-site. **(c)** Alternative 5′ splice-site. **(d)** Mutually exclusive exons. **(e)** Intron retention.

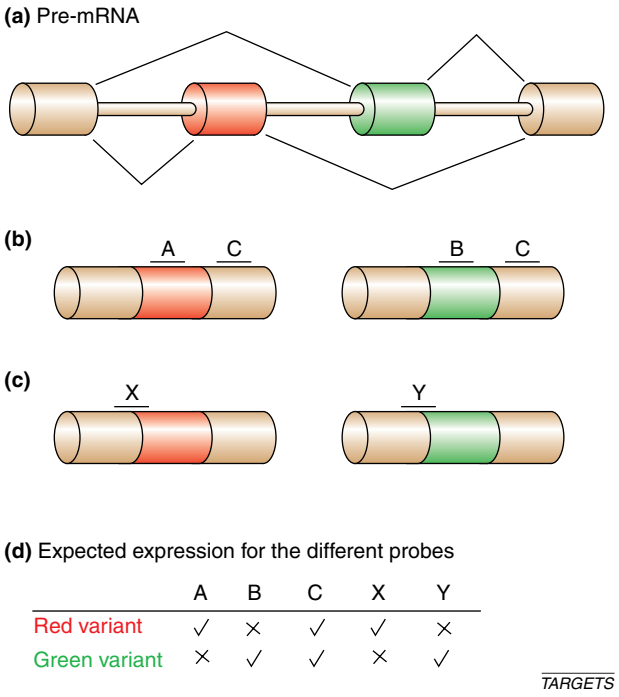


**Figure 2.** Two possible ways of designing microarrays that report on alternative splicing. **(a)** Schematic representation of a pre-mRNA with two splicing options. **(b)** Designing probes unique to regions from specific variants. **(c)** Designing probes that connect junctions between two exons. **(d)** The signal expected by each of the probes.

if alternative splicing is not taken into consideration. In the next few paragraphs we will describe a few of the main techniques for target identification, target validation and drug design, and explain how they can be enhanced through awareness of alternative splicing.

## DNA microarrays

DNA microarrays (chips) are the method of choice for comparing the transcriptomes of different cell types or cells under different conditions, as well as for toxicity analyses. In this method, short probes, each representing a gene, are printed on a slide. The cDNA of the tested tissue, fluorescently labeled, is then hybridized to the microarray, and the level of the resulting signal should reflect the relative mRNA expression level of the gene. To design a DNA microarray that will effectively report on the transcriptional levels of genes, however, alternative splicing patterns must be taken into account.

The possible consequences of designing probes for an alternatively spliced gene without considering alternative splicing can be inferred from Figure 2. Taking the probe from the region common to both variants (Fig. 2b, probe C) will produce a combined signal for both transcripts, which will be especially misleading if the variants have different expression profiles (as often occurs). If the probe is unique for only one of the variants, however, the expression of the other will not be measured and the information about the expression of the gene will be incomplete (Fig. 2b, probe B). In such a case one of the variants will be left out of our view of the transcriptome and therefore go undetected as a potential marker or target.

Currently, the commonly used DNA microarray technologies do not consider alternative splicing, or do so only in a very limited way. In most DNA microarray design technologies, the probe is taken from the 3′ end of the gene. Such a design strategy combines the signals from all the splice variants occurring within the gene to one signal. In addition, it ignores variants that result from alternative polyadenylation, a common transcriptomic phenomenon [17,18]. As

about half of all human genes are alternatively spliced, the spectrum of transcripts that could currently be detected by DNA chips is much smaller than the actual spectrum.

Pioneering technologies addressing these problems have emerged in recent years. These technologies use alternative splicing knowledge to design DNA chips that report on splice variants. In the approach presented by Hu *et al.* [19], each variant is represented by a specific set of probes, so the expression level of separate variants can be measured directly (Fig. 2b). Using this approach, the authors were able to validate computer-predicted rat alternative splicing forms [19]. Implementing these principles, the widely used U133 array of Affymetrix was designed to contain around 6000 alternatively spliced variants of known genes (www.affymetrix.com), but this, as described above, is only a small fraction of the existing variants.

Other technologies, such as the Bead Array RASL technology [20] and the splicing-specific microarrays used by Clark *et al.* [21], detect splice variants using probes spanning splice junctions, that is, boundaries between two exons (Fig. 2c). Such probes will only detect variants in which the two exons are adjacent, and therefore can measure expression levels of specific variants.

If, on the other hand, one is interested in the complete expression level of a gene, arrays could be designed that will detect all the variants of a specific gene with one probe. Such an approach is used in the OligoLibraries™ (of Sigma-Genosys and Compugen), which are based on probe selection using Compugen's LEADS™ [22].

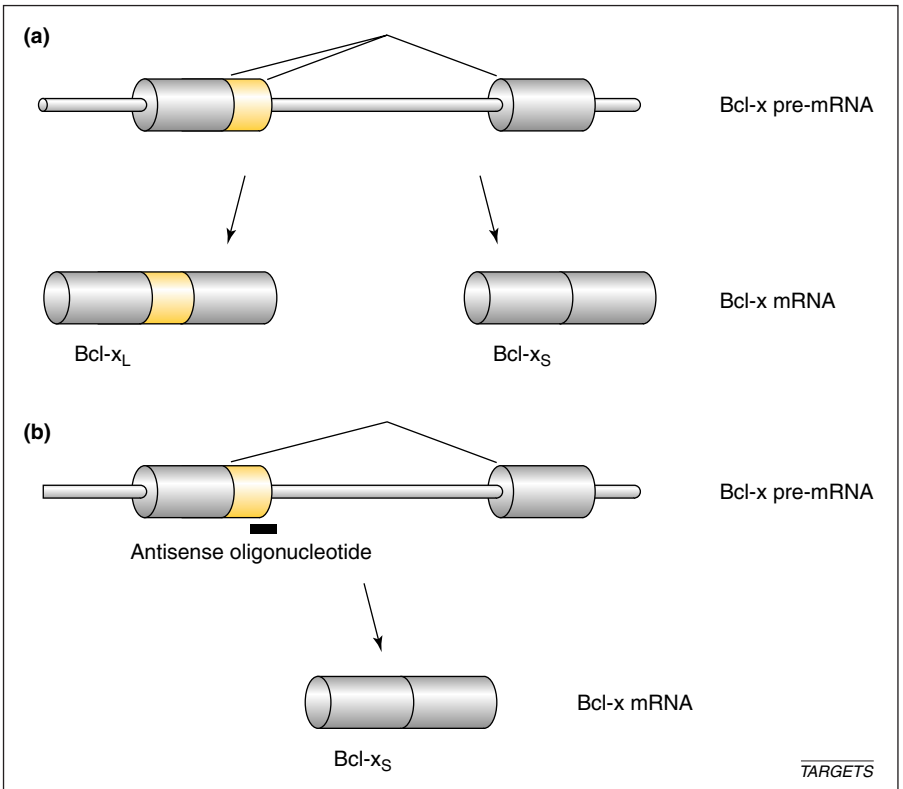


**Figure 3.** Alternative splicing of Bcl-x pre-mRNA is modified by antisense oligonucleotide. **(a)** Two possible splice variants of Bcl-x. The long variant (Bcl-$x_L$) is the anti-apoptotic variant; the short variant (Bcl-$x_S$) is pro-apoptotic. **(b)** Addition of antisense oligonucleotide matching the unique 5′ splice site of the long variant induces a shift of splicing from Bcl-$x_L$ to Bcl-$x_S$. This results in a significant apoptosis phenotype in PC-3 prostate cancer cells (adapted from [24]).

## Antisense and RNA interference

Antisense-mediated silencing and RNA interference (RNAi) are two commonly used methods for knock-down of specific genes. In the antisense approach, short antisense oligonucleotides complementary to a region in the mRNA molecule that needs to be silenced are introduced to the cell. RNAi is a powerful new technique in which short interfering RNA (siRNA) causes degradation of the mRNA corresponding to the siRNA sequence (reviewed in [23]).

Here, too, understanding the alternative splicing pattern of the gene targeted for silencing is essential for effective results. If a complete shutdown of the gene is required, it is important that the region selected to be represented as an antisense oligonucleotide or siRNA will be common to all splice variants. If, on the other hand, the silencing of only one splice variant is desired, the oligonucleotides/siRNAs should be selected from the region unique to this variant.

Antisense oligonucleotides can also be designed to be complementary to the splice sites of the targeted variant, thus 'guiding' the spliceosome to the 'right' variant. Indeed, such an approach was implemented to down-regulate a splice variant of the *Bcl-x* gene, an apoptotic regulator of the cell [24] (Fig. 3). This gene has two splice variants, Bcl-$x_L$ (long isoform) and Bcl-$x_S$ (short isoform), with opposite functions: the short isoform is pro-apoptotic and the long one is anti-apoptotic. Bcl-$x_L$ is highly expressed in many types of cancer, including multiple myeloma, small cell lung carcinoma and breast cancer. Specifically, 100% of prostate tumors express Bcl-x, presumably the long variant. Using antisense oligonucleotides against the 5′ splice site of the long, anti-apoptotic variant, Mercatante *et al.* managed to shift the splicing pattern towards the short, pro-apoptotic isoform and to induce apoptosis in a prostate-cancer cell line [24].

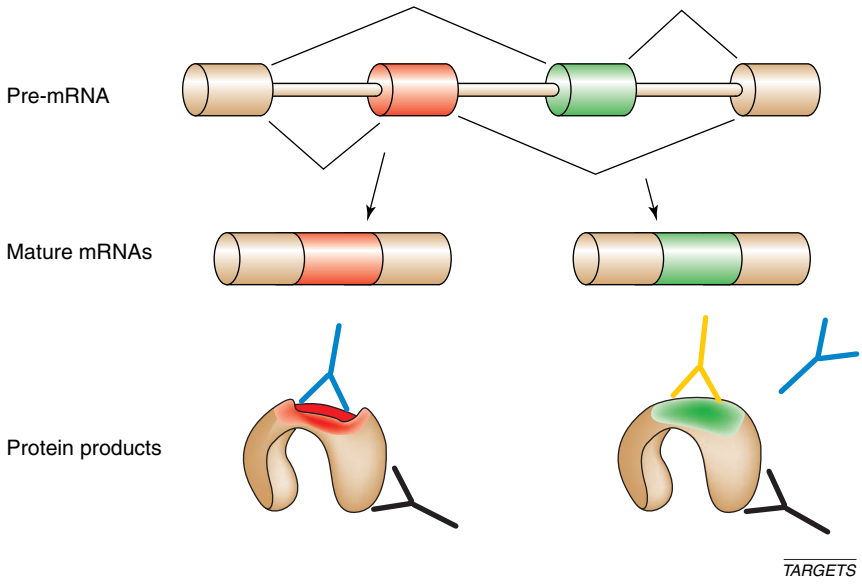


**Figure 4.** Variant-specific antibodies can distinguish between different protein isoforms. Schematic representation of a gene with two splice variants resulting from the 'mutually exclusive exons' pattern of alternative splicing. Each of the variants is translated to a protein isoform. Regions common to both protein isoforms are beige; isoform-specific regions are red and green. The black antibody binds the common region, and will therefore bind both variants. The blue and yellow antibodies are isoform-specific.

Antisense oligonucleotides can be designed to inhibit abnormal, disease-causing splicing. Such an approach was used to correct an aberrant splicing of the cystic fibrosis transmembrane conductance regulator (CFTR) gene [25]. This approach was also employed to alter the alternative splicing of the *Tau* gene that causes frontotemporal dementia and parkinsonism linked to chromosome 17 [26]. Similarly, Celotto and Graveley demonstrated that exon-specific RNAi can be used to selectively degrade specific alternatively spliced mRNA isoforms containing exons 4.1 or 4.4 of the *Dscam* gene in cultured *Drosophila* cells [27].

Antisense RNA complementary to a unique splice variant could also be used to induce PKR-mediated cell death. PKR is a kinase that is activated in the presence of double-stranded RNA longer than 30 base-pairs and, when activated, it induces apoptosis and inhibition of protein synthesis. Shir and Levitzki demonstrated that they could recruit PKR for the inhibition of glioma growth by using antisense RNA complementary to fragments flanking a glioma-specific deletion [28]. As some splice variants are restricted to specific tumors, future implementation of this method could target these splice variants as a means to eliminate the tumors.

## Other technologies: antibodies, mass spectrometry and knockout

Alternative splicing information is useful at many other stages of the drug discovery process. For example, when selecting an antibody for a desired therapeutic role, alternative splicing must be considered. If, for example, one wants to eliminate totally a certain protein, the antibody should be designed against the region common to all splice variants of it; otherwise, the antibody will miss some of the protein isoforms (Fig. 4). Alternatively, when only one of the isoforms is disease-specific, an antibody that binds only the isoform-specific part of the protein should be selected. In both these cases, a prior knowledge of alternative splicing is necessary. In recent years, many isoform-specific monoclonal antibodies have been used by academic investigators for basic research; however, no variant-specific drugs have yet been developed.

Knowledge of alternative splicing is also important when designing knockout mice experiments. For example, the mouse knockout of the estrogen receptor α (ERα) was used for several years as a model to study the function of the gene. However, it turned out that not all the gene is shut down and a residual binding activity remains in uterine tissue. A recent study showed that this activity resulted from a splice variant of the ERα that was not knocked out [29]. Thus, when trying to shut down the activity of a gene, the knockout cassette should be inserted into an exon common to all variants. It is also possible to design knockout mice for specific isoforms.

Another example concerns mass spectrometry, which is a technique enabling the identification of a protein by comparing the mass of peptides derived from its proteolytic cleavage to a database of known protein sequences. Usually a small number of identified peptides is sufficient for protein identification. Without knowledge of alternative splicing, mass spectrometry results may be incorrectly interpreted, as some peptides are common to several protein isoforms.

## Alternative splicing prediction tools

The finding that alternative splicing could be deduced from alignments of expressed sequence tags (ESTs) and full-length mRNAs was a major breakthrough in alternative splicing research [30]. An EST represents a piece of an mRNA, and is the result of sequencing part of a cDNA clone that has been generated from an mRNA [31]. The largest public database of ESTs is dbEST [32], a division of GenBank that

currently contains >16 million sequences, including >5 million ESTs from human. There are also private EST databases, such as the one produced by Incyte Genomics (www.incyte.com), which also contain millions of ESTs.

Using multiple alignment algorithms, the millions of available ESTs and cDNAs can be utilized to identify alternative splicing events (Fig. 5). Such analyses have led to the estimate that 35–59% of all human genes are alternatively spliced [7–11]. Several resources provide specific alternative splicing analysis for genes in human and other organisms (Table 1).

Large-scale analysis of EST data involves a considerable computational complexity. First, the alignment of this huge amount of sequence data requires powerful computers running accelerating algorithms. Just to give the sense of it, if an alignment of one EST to the genome takes ten seconds with a public query in NCBI, without using custom algorithms and powerful hardware it would take >1.5 years to align the five million human ESTs to the human genome.

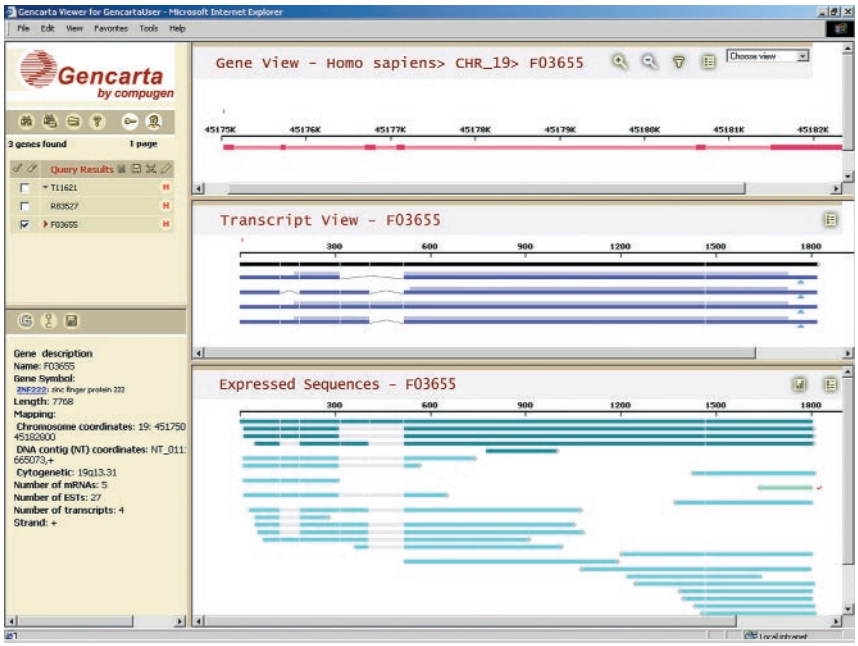


**Figure 5.** Deduction on alternative splicing using alignments of expressed sequences. Lower right panel graphically represents alignment of expressed sequences. ESTs are blue; full-length mRNAs are dark green. Middle right panel shows four distinct splice isoforms (transcripts) deduced from the alignment. Deduced mRNA transcript are purple; open reading frames are light purple. Upper right panel shows the genomic organization of this gene. Exons are dark red boxes; introns are light red. The figure presented is a snapshot of the Gencarta™ transcriptome viewer; other tools, such as the UCSC genome browser and the Ensembl genome browser, are also available (see Table 1).

Another obstacle is the fact that EST datasets intrinsically contain countless artefacts that complicate the accurate prediction of splice variants [33,34], including bad sequence quality, vector contamination, chimeric sequences and so on. Some kinds of contamination are easily detected, whereas others, such as pre-mRNA contamination and genomic contamination, are more subtle. As the artefacts often result in spurious alternative splicing predictions, these predictions must be viewed with caution. Recently, new methods have been developed for screening such artefacts [34,35] that use statistical features of confirmed

**Table 1. Current online resources for alternative splicing analysis**

| Genomic viewers | Website | Description |
|---|---|---|
| UCSC genome browser | http://genome.ucsc.edu/ | All three genomic viewers contain a graphical interface to results of alignments of expressed sequences to the genome, and visually show splice variants inferred from the alignments. |
| Ensembl genome browser | http://www.ensembl.org/ | |
| Gencarta™ | http://www.gencarta.com | |
| **Alternative splicing databases** | **Website** | **Description** |
| ASAP | http://www.bioinformatics.ucla.edu/ASAP/ | Based on genome-wide analyses of alternative splicing in human |
| HASDB | http://www.bioinformatics.ucla.edu/~splice/HASDB/ | Human alternative splice variants predicted from EST data |
| AsMamDB | http://166.111.30.65/ASMAMDB.html | Human, mouse and rat alternative splicing database |
| Alternative Splicing Database | http://cgsigma.cshl.org/new_alt_exon_db2/ | Alternative splicing evidence collected from the literature |

alternative splicing and of EST libraries. Another verification method is to check the conservation of alternative splicing between species [9,36].

Despite the possible difficulties and artefacts, ESTs are still the major tool for large-scale analyses of alternative splicing, and are currently the best data source for alternative splicing information. However, it is also possible to use other tools, such as the DATAS technology by ExonHit (www.exonhit.com). In this technology, alternatively spliced exons are identified by an experimental comparison between mRNAs from two different tissues. SAGE (serial analysis of gene expression) is another technology, in which the expression levels of short sequence tags representing 3′ ends of different mRNAs are measured. This method is much cheaper than the EST method, as it only requires the sequencing of 10 base-pairs per transcript, but it is less accurate and can mainly detect alternative poly-adenylation rather than alternative splicing.

## Conclusions

Clearly, knowledge of alternative splicing is crucial to the key drug discovery processes. We have reported that technologies enabling the usage of alternative splicing information are emerging, but are still far from being mature. More ways of implementing alternative splicing knowledge, such as in the fields of rational structure-based drug design, chemical screening, pharmacogenomics, pathways and toxicity analyses, will probably emerge in the future. This new window to the transcriptome and the proteome adds an additional dimension of accuracy and reduces the costly levels of uncertainty embedded in the process. Already there are several biotechnology companies (among them are Compugen and ExonHit) that base their drug discovery strategy on the new knowledge of alternative splicing. The next few years will probably see many more drug-discovery programs that will embrace this new knowledge.

## Acknowledgements

We thank Galit Rotman for commenting on earlier versions of this manuscript.

## References

1 Waterston, R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562
2 Will, C.L. and Luhrmann, R. (2001) Spliceosomal UsnRNP biogenesis, structure and function. *Curr. Opin. Cell Biol.* 13, 290–301
3 Muller, S. *et al.* (1998) A supraspliceosome model for large nuclear ribonucleoprotein particles based on mass determinations by scanning transmission electron microscopy. *J. Mol. Biol.* 283, 383–394
4 Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17, 100–107
5 Maniatis, T. and Tasic, B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418, 236–243
6 Schmucker, D. *et al.* (2000) Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* 101, 671–684
7 Brett, D. *et al.* (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* 474, 83–86
8 Croft, L. *et al.* (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat. Genet.* 24, 340–341
9 Kan, Z. *et al.* (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.* 11, 889–900
10 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
11 Mironov, A.A. *et al.* (1999) Frequent alternative splicing of human genes. *Genome Res.* 9, 1288–1293
12 Okazaki, Y. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563–573
13 Wu, J.Y. *et al.* (2003) Alternative pre-mRNA splicing and regulation of programmed cell death. *Prog. Mol. Subcell. Biol.* 31, 153–185
14 Krawczak, M. *et al.* (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.* 90, 41–54
15 Krawczak, M. and Cooper, D.N. (1997) The human gene mutation database. *Trends Genet.* 13, 121–122
16 Koslowski, M. *et al.* (2002) Multiple splice variants of lactate dehydrogenase C selectively expressed in human cancer. *Cancer Res.* 62, 6750–6755
17 Beaudoing, E. and Gautheret, D. (2001) Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res.* 11, 1520–1526
18 Gautheret, D. *et al.* (1998) Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res.* 8, 524–530
19 Hu, G.K. *et al.* (2001) Predicting splice variant from DNA chip expression data. *Genome Res.* 11, 1237–1245
20 Yeakley, J.M. *et al.* (2002) Profiling alternative splicing on fiber-optic arrays. *Nat. Biotechnol.* 20, 353–358
21 Clark, T.A. *et al.* (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* 296, 907–910
22 Shoshan, A. *et al.* (2001) Designing oligo libraries taking alternative splicing into account. In *Proceedings of SPIE: Microarrays: Optical Technologies and Informatics* (Vol. 4266) (Bittner, M.L. *et al.*, eds.), pp. 86–95
23 Opalinska, J.B. and Gewirtz, A.M. (2002) Nucleic-acid therapeutics: basic principles and recent applications. *Nat. Rev. Drug Discov.* 1, 503–514
24 Mercatante, D.R. *et al.* (2001) Modification of alternative splicing of Bcl-x pre-mRNA in prostate and breast cancer cells. Analysis of apoptosis and cell death. *J. Biol. Chem.* 276, 16411–16417
25 Friedman, K.J. *et al.* (1999) Correction of aberrant splicing of the cystic fibrosis transmembrane conductance regulator (CFTR) gene by antisense oligonucleotides. *J. Biol. Chem.* 274, 36193–36199
26 Kalbfuss, B. *et al.* (2001) Correction of alternative splicing of tau in frontotemporal dementia and parkinsonism linked to chromosome 17. *J. Biol. Chem.* 276, 42986–42993
27 Celotto, A.M. and Graveley, B.R. (2002) Exon-specific RNAi: a tool for dissecting the functional relevance of alternative splicing. *RNA* 8, 718–724
28 Shir, A. and Levitzki, A. (2002) Inhibition of glioma growth by tumor-specific activation of double-stranded RNA-dependent protein kinase PKR. *Nat. Biotechnol.* 20, 895–900
29 Kos, M. *et al.* (2002) Down but not out? A novel protein isoform of the estrogen receptor α is expressed in the estrogen receptor α knockout mouse. *J. Mol. Endocrinol.* 29, 281–286
30 Black, D.L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* 103, 367–370
31 Adams, M.D. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651–1656
32 Boguski, M.S. *et al.* (1993) dbEST–database for 'expressed sequence tags'. *Nat. Genet.* 4, 332–333
33 Modrek, B. and Lee, C. (2002) A genomic view of alternative splicing. *Nat. Genet.* 30, 13–19
34 Sorek, R. and Safer, H.M. (2003) A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.* 31, 1067–1074
35 Kan, Z. *et al.* (2002) Selecting for functional alternative splices in ESTs. *Genome Res.* 12, 1837–1845
36 Sorek, R. and Ast, G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* (in press)