Luv Kashyap* and Parul Tripathi**

*Department of Biochemistry, Aligarh Muslim University, Aligarh, India
** Immunology Group, International Centre for Genetic Engineering and Biotechnology, New Delhi, India

Alternative Splicing
How one gene can make many proteins

Address for correspondence:

Dr. Parul Tripathi,
Research Scientist
Immunology Group
International Centre for Genetic Engineering and Biotechnology (ICGEB),
New Delhi, India
E-mail: parultripathi@rediffmail.com

Key words:

human genome, mRNA, bioinformatics, introns, exons, mutations

Short descriptive title:

Alternative Splicing "One gene many proteins"

Humans have only twice as many genes as the simple fruit fly. How can this be? After all, we have brains that allow us to count genes while their main occupation is seeking out a nice banana.
Alternative splicing explains the compact living of our genetic information as well as the mechanisms behind several human diseases.

The sequencing of the human genome (1) has raised important questions about the nature of genomic complexity. Scientists thought that the complex DNA of a human was made up by perhaps as many as 150,000 different genes. The approximation was based on the number of different transcripts, mRNAs, which had been found in humans, assuming that there should be one gene for each mRNA. When all the sequencing was done, the result was merely 32,000 human genes. (Later results suggest that the number is even lower, less than 25,000.) In comparison, the fruit fly *Drosophila* has 14,000 genes and the nematode *Caenorhabditis* elegans has 19,000. The report of only 32,000 human genes came as a surprise (2). The number of human expressed-sequence (mRNA) forms was much higher than the number of genes. How can the much greater size and complexity of humans be encoded in only twice the number of genes required by a fly? One way to explain this paradox is to point out that the number of possible proteins from the genome can far exceed the number of genes if a large percentage of the genes have the ability to encode multiple proteins. This expansion of the proteome can be accomplished through alternative precursor messenger RNA (pre mRNA) splicing.

What is alternative splicing?

Alternative splicing is a process which allows the production of a variety of different proteins from one gene only (3, 4) (**Figure 1**). Most genes in eukaryotic genomes consist of exons and introns. After transcription, introns need to be removed from the pre-mRNA by a step called splicing. Sometimes an exon can be either included or excluded from the final transcripts, or there can be two splice sites at one end of an exon that are recognized by the spliceosome (the complex which carries the splicing reaction). This makes it possible to produce multiple transcripts by "alternative splicing". The joining of different splice sites allows individual genes to express multiple mRNAs that encode proteins with diverse and even antagonistic functions (5). Through this mechanism, the information stored in the genes of complex organisms can be edited in a variety of ways, making it possible for a single gene to specify two or more distinct proteins. It has been proposed that for eukaryotes it was a very important step towards higher efficiency, because information can be stored much more economically.

Understanding the functional impact of alternative splicing, the regulatory mechanisms that govern RNA splicing, and the role of alternative splicing in genome evolution, has become one of the most challenging and exciting tasks for genomics and bioinformatics in the post-genomic era. The types of alternative splicing alteration that have been observed include exon skipping; intron retention, use of alternative splice donor or acceptor site and many more.
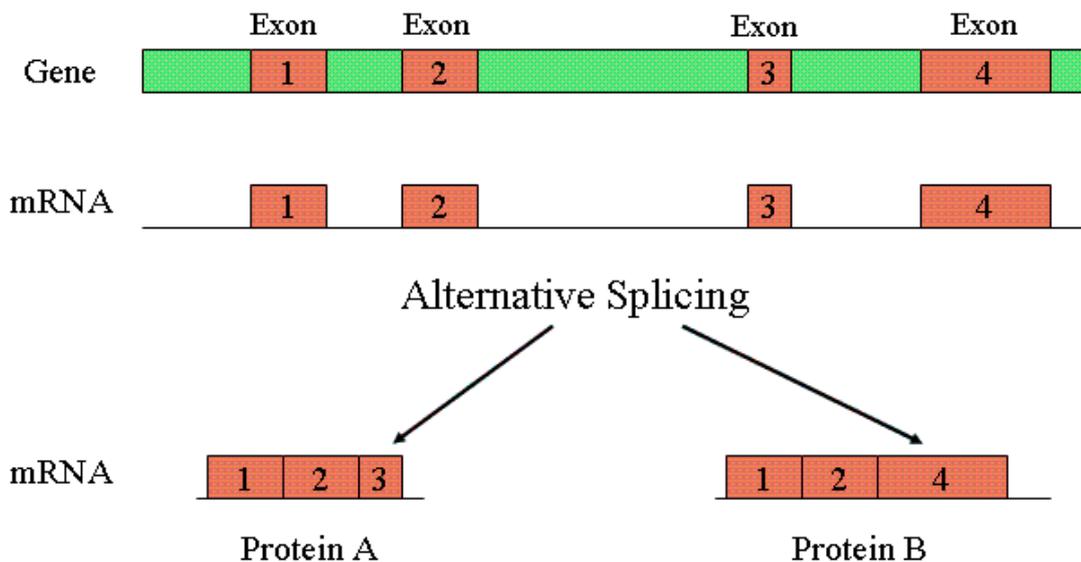
Figure 1: Demonstration of alternative splicing: *In DNA, the genetic information that includes the code for making a protein is located in fragments (exons, red boxes), which are interrupted by non-coding fragments (introns, green boxes). By the process of alternative splicing, the introns are removed and the exons spliced together in different combinations, generating different messenger RNAs (mRNA) that are decoded (translated) into distinct proteins.*

A surprisingly common process

Alternative splicing has long been regarded as a rather rare event in eukaryotic genomes. For example, it was estimated to occur in less than 5% of human genes. However, recent analyses of vast amounts of transcript data in human and other organisms suggest that alternative splicing is widespread in mammalian genomes. In humans, it is estimated that alternative splicing occurs in more than 60% of genes. The ***one gene, one protein*** central dogma, a governing theory of modern molecular biology which has profound impact on biologists' thinking, therefore has to be revisited. As a result, our view towards many biological processes, such as protein interaction and gene expression, has to be adjusted.

Alternative splicing is a powerful means of enhancing protein diversity. It is estimated that over 70% of all genes are alternatively spliced as a means for producing functionally diverse proteins from a single gene. In fact, the majority of metazoan genes encode pre-mRNAs that are alternatively spliced to produce anywhere from two to tens of thousands of mRNA isoforms. Alternative splicing is found extensively in all higher eukaryotes, from organisms like *Caenorhabditis elegans,* (6, 7), *Drosophila* (8), mouse (9) to humans, emphasizing the importance of alternative splicing throughout evolution. One of the most dramatic examples of alternative splicing is the *Dscam* gene in *Drosophilia* (8). This single gene contains some 116 exons of which 17 are retained in the final mRNA. Some exons are always included; others are selected from an array. Theoretically this system is able to produce 38,016 different proteins. And, in fact, over 18,000 different ones have been found in *Drosophila*.

The Dscam proteins are involved in guiding neurons to their proper destination and probably in recognition and phagocytosis of invading bacteria. Whether a particular segment of RNA will be retained as an exon or excised as an intron depends on the circumstances.

Regulating genes

Alternative splicing plays an important role in regulating gene expression. In the last decade it has been shown that alternative splicing determines binding properties, intracellular localization, enzymatic activity, protein stability and post-translational modifications of a large number of proteins. Many proteins are comprised of several domains, or modules, that serve a particular function. For example, one domain may help the protein bind to another protein, while another domain gives the protein enzymatic activity. In the genome, domains correspond to exons. By alternative splicing exons, i.e. protein domains, can be mixed and matched, altering the nature of the protein. By regulating which splice patterns occur in which tissue, an organism fine-tunes the action of a single gene so that it can perform many different roles.

The consequences of alternative splicing fall into two major categories: **protein-level alterations** and **transcript-level modifications**. On the *protein level*, alternative splicing generates splice variants that give rise to different protein products. For example, a shortened protein product due to a frame shift introduced by an alternate exon or a protein product with a different functional domain due to the inclusion of a specific exon from a mutually exclusive group of exons. On the *transcript level*, alternative splicing produces splice variants that have different translation or stability profiles. For example, a transcript with a longer lifespan would prolong the availability of the corresponding protein. Thus, it plays an extremely important role in expanding protein diversity and might therefore partially explain the apparent discrepancy between gene number and organism complexity (1).

Alternative splicing and human disease

Alternative splicing events have been implicated in a large number of human pathologies, including neurodegenerative, cardiovascular, respiratory and metabolic diseases, as well as cancer and Alzheimer's disease, and are promising targets for therapeutic intervention. In tissues with complex activities, like the brain, there is a demand for a large number of proteins to perform different functions. Any defect in alternative splicing can cause severe problems for the cell since the protein composition is altered, and many neurological diseases are actually caused by defects in alternative splicing. These defects can be divided into two major groups: Primary and secondary splicing defects (17).

Primary splicing defects

In primary splicing defects, sequences on the pre-mRNA which are important for splicing are mutated. Consequently, correct splicing cannot occur (**Figure 2**). Many point mutations causing human disease are considered to cause defects in pre-mRNA splicing. Examples are ataxia-telangiectasia and neurofibromatosis. Half of the patients suffering from those diseases carry mutations that effect pre-mRNA splicing (18, 19). Another example for primary splicing disorders is found in fronto-temporal dementia patients. This syndrome is an autosomal dominant disorder linked to chromosome 17. Splicing defects result in altered transcripts of the microtubule associated protein tau (MAPT) which is enriched in axons of mature and growing neurons.

Altered splicing of MAPT has also been implicated in the pathology of other diseases such as Alzheimer's disease, myotonic dystrophy, gliopathy and spinal cord degeneration. Duchenne muscular dystrophy (DMD) is an effect of mutations in the *dystrophin* gene. Mutations cause inappropriate splicing and production of incorrect mRNAs. A different mutation in the *dystrophin* pre-mRNA is better tolerated and results in a milder variant of DMD, Becker's muscular dystrophy (BMD) (17).
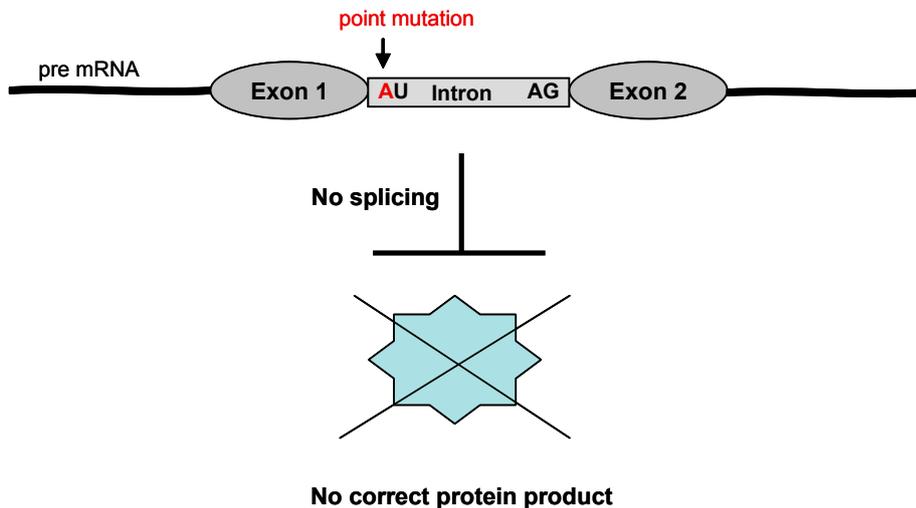
A
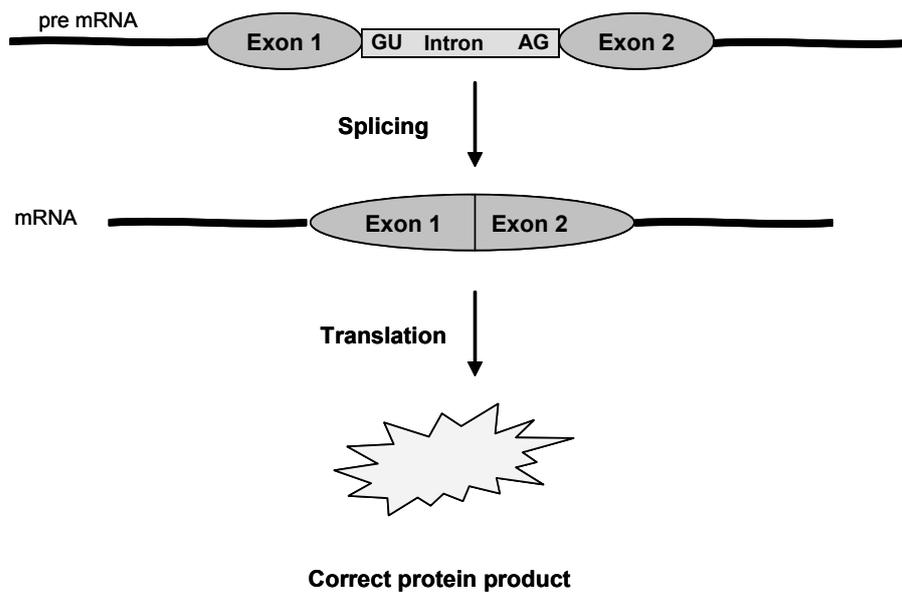
B

Figure 2 (A): *Example of correct splicing. Exon 1 and Exon 2 are the parts of the pre-mRNA which carry the coding information in the mRNA for the production of a protein. The intron is the non-coding sequence in between which is removed; Exon 1 and Exon 2 are spliced together. After splicing, the mature mRNA is used for the production of a protein.* **Figure 2 (B):** *Example of a primary splicing defect. The invariant sequence GU, which defines the beginning of the intron, is mutated. The splice site is incorrect (AU instead of GU for example as a result of a point mutation in the sequence) and cannot be used any more. The intron is not spliced out and no correct protein is produced.*

www.gu.se

Secondary splicing defects

In secondary splicing disorders, a regulatory factor which is essential for the process of splicing, is mutated and disturbs splicing activity. Differences in the regulation of splicing activator and/or repressor proteins can have an effect on the alternative splicing pathway. Depending on which splicing regulator is affected, the alternative splicing pattern of particular pre-mRNAs is changed. In some cases, this can lead to severe disturbances whereas in others, milder disease patterns are seen. One such example is shown in **Figure 3**. Two possible splice sites (AG) are accessible to give two different mature mRNAs and therefore two different protein products (Protein Z and Protein Y). The splicing regulator (X) binds to a splice site which is located within a coding region of the pre-mRNA. This splice site is now no longer accessible to the splicing machinery and cannot be used for splicing Factor X can be altered (for example by a point mutation) and therefore be impaired in binding to Exon 2.
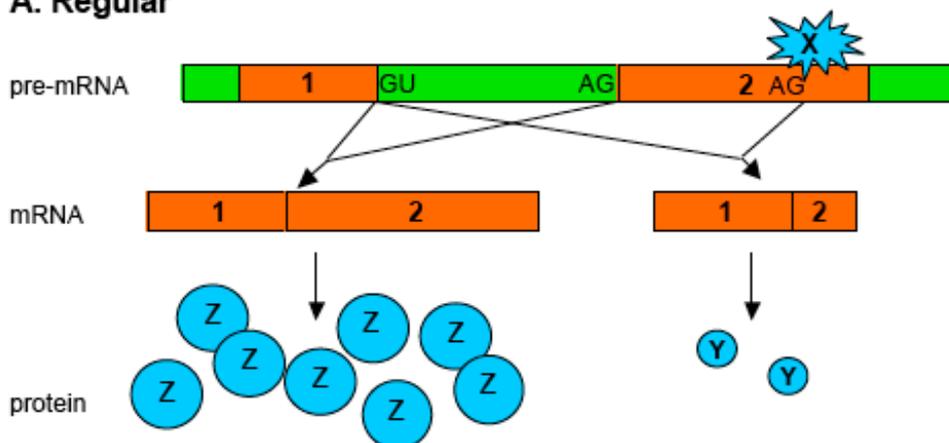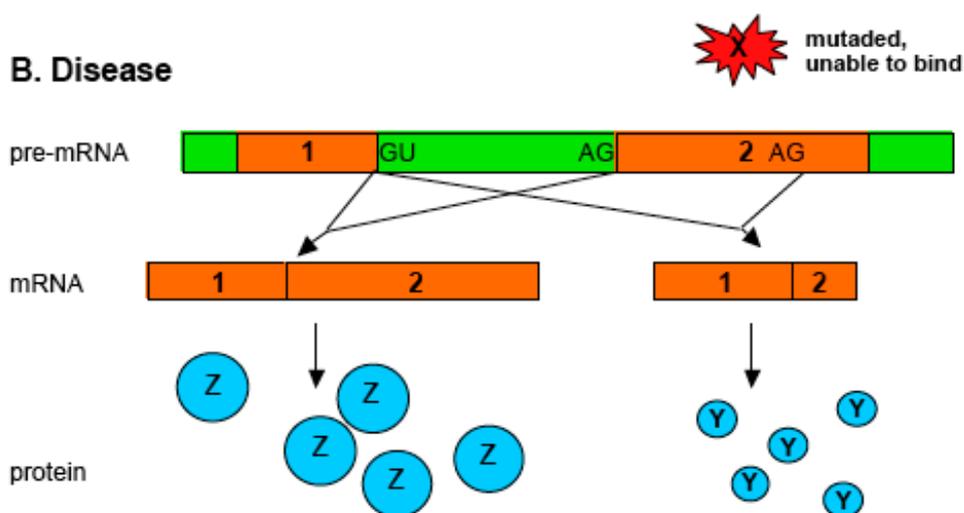
Figure 3: *Secondary splicing defects. In this pre-mRNA two different splice sites (AG) can be used for splicing. The two resulting products after splicing and translation (Protein Z and Protein Y) are different depending on the splice site which was used. In the normal situation (panel A) one splice site is partly blocked by a protein factor X. Only the second site at the end of the intron is easily accessible and therefore one of the proteins, Protein Z, is produced in larger amounts than Protein Y. After a mutation (panel B), the splicing regulator X cannot bind to the pre-mRNA any more. Both AG splice sites are accessible and can be used for splicing. When the percentage of Protein Z is reduced and Protein Y levels increase, this causes disease.*

One example for a secondary splicing defect is the Prader Willi Syndrome (PWS), a genetic disorder characterized by intellectual and behavioural disturbances. The disease may arise from the absence of a small RNA molecule which normally regulates pre-mRNA splicing of the serotonin receptor RNA. The enzyme acetylcholine esterase (AChE) is a combinatorial series of proteins which is produced via alternative splicing events. Three alternatively spliced isoforms have been identified which are involved in the progression of neurodegenerative disorders like Alzheimer's disease and Parkinson's disease. An imbalance of splicing regulators can cause modified alternative splicing of AChE pre-mRNA and lead to an excess of AChE-R, an alternatively spliced isoform of AChE, in patients suffering from Alzheimer's disease. In the case of Parkinson's disease injury and psychological stress (leading causes of the disease) are associated with an imbalance in AChE alternative splicing and overexpression of the alternate isoforms AChE-R (20).

Future prospects

Alternative splicing has not only emerged as a key step in controlling mammalian gene expression, but is also a major source of diversity in the human proteome.
Today "expressed sequence tags" or "microarray analysis" (see Box 1) are the preferred methods for large-scale detection of alternative splicing (12-15). The methods do have limitations and it is difficult to detect all alternative splice events. Newer and more effecient methods are being developed or are in process of being developed which could detect all possible spliced transcripts of a gene. So, the estimated prevalence of alternative splicing changes every year. The higher prevalence of alternative splicing makes it an imperative to produce a comprehensive catalog of all splice variants and develop a better understanding of its origin, function, and regulation (11).

When control of alternative splicing is disrupted, the result can be a failure to meet cellular and tissue requirements resulting in dysfunction and disease. Investigations into how the misregulation of alternative splicing causes disease complements investigations of normal regulatory processes and enhances our understanding of regulatory mechanisms in general. Ultimately, an understanding of how alternative splicing is altered in disease will facilitate strategies directed at reversing or circumventing misregulated splicing events.

> **Box 1**
>
> EST (= expressed sequence tags) is a short sub-sequence of a transcribed spliced nucleotide sequence. They may be used to identify gene transcripts, and are instrumental in gene discovery and gene sequence determination
>
> EST sequences enhance the effectiveness of molecular studies, especially for gene expression profiling and the analysis of genes involved in virulence and infectivity.
>
> A DNA microarray (also commonly known as gene or genome chip, DNA chip, or gene array) is a collection of microscopic DNA spots, commonly representing single genes, arrayed on a solid surface by covalent attachment to a chemical matrix.
> From Wikipedia)

Some questions for students:

1. How do we account for complexity of higher organisms like humans with just the gene pool not substantially higher than the number of protein coding genes in the genome of the nematode (19 000), and less than in rice (38 000–40 000) ?

2. What is the Central dogma of Molecular Biology? Why is this true any longer?

3. What are the differences between primary and secondary splicing defects?

4. Search the web for some of the diseases caused by failure in alternative splicing and how they can be recognized.

5. Find out more about microarray technique and how it can be used when studying alternative splicing!

Some interesting websites to explore

http://www.alternative-splicing.org/org/index.html

http://en.wikipedia.org/wiki/Alternative_splicing

http://www.rnabioinformatics.com/2004/10/intro_what_is_alternative_spli.html

http://scienceandreason.blogspot.com/2006/11/alternative-splicing.html

References

International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature* 409 860-921

Venter, J.C. Adams, MD. Myers, EW. Li, PW. Mural, RJ. *et al* (2001). The sequence of the human genome. *Science* 291 1304-1351.

Maniatis, T. and Tasic, B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418 236–243.

Matlin, A.J. Clark, F. and Smith, CWJ. (2005) Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* 6 386–398.

Faustino, NA. and Cooper, TA. (2003) Pre-mRNA splicing and human disease. *Genes Dev* 17 419–437.

Tabish, M. Clegg, RA. Rees, HH. and Fisher, MJ. (1999) Organization and alternative splicing of the *Caenorhabditis elegans* cAMP-dependent protein kinase catalytic-subunit gene (*kin-1*). *Biochem J* 339 209-216.

Park, YS. S. Kim, Y. Shin, BC. and Cho, NJ. (2003) Alternative splicing of the muscarinic acetylcholine receptor GAR-3 in Caenorhabditis elegans. *Biochem Biophys Res Commun* 308 961-965.

Celotto, AM and Graveley, BR. (2001) Alternative splicing of the Drosophila Dscam pre-mRNA is both temporally and spatially regulated. *Genetics* **159** 599–608.

Waterston, RH. Lindblad-Toh, K. Birney, E. *et al.,* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420 520–562

Resch, A. Xing, Y. Modrek, B. Gorlick, M. Riley, R. and Lee, C. (2004) Assessing the impact of alternative splicing on domain interactions in the human proteome. *J Proteome Res* 3 76–83.

Johnson, J. Castle, J. Garrett-Engele, P. *et al* (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 30**2** 2141-2144.

Kashyap, L. and Tabish, M (2007c) Alternatively spliced isoforms encoded by cadherin genes from *C. elegans* genome. Bioinformation., 2, 50-56.

Kashyap, L. Tabish, M., Ganesh, S. *et al.,* (2007a) Computational and molecular characterization of multiple isoforms of *lfe-2* gene in nematode *C. elegans* Bioinformation., 2, 17-21.

Kashyap, L. Tabish, M., Ganesh, S. *et al.,* (2007b) Identification and comparative analysis of novel alternatively spliced transcripts of RhoGEF domain encoding gene in *C. elegans* and *C. briggsae*, Bioinformation., 2, 43-49.

Kashyap, L. and Tabish, M. (2006). Comparative analysis of various gene finders specific to *Caenorhabditis elegans* genome. Bioinformation., 1, 203-207.

Xia, H., Bi, J., and Li, Y. (2006) Identification of alternative 5'/3' splice sites based on the mechanism of splice site competition. Nucleic Acids Res. 34, 6305-6313.

Cáceres JF, Kornblihtt, AR. (2002). Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet.,18, 186-93.

Ars E, Serra E, Garcia J, Kruyer H, Gaona A, Lazaro C, Estivill X. (2000) Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum Mol Genet., 9,237-247.

Teraoka SN, Telatar M, Becker-Catania S, Liang T, Onengut S, Tolun A, Chessa L, Sanal O, Bernatowska E, Gatti RA, Concannon P. (1999) Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. Am J Hum Genet., 64,1617-1631.

Meshorer E, Soreq H. (2006) Virtues and woes of AChE alternative splicing in stress-related neuropathologies. Trends Neurosci., 29,216-224.