Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Views & Challenges**

Alternative splicing: a paradoxical qudo in eukaryotic genomes

Luv Kashyap[1] and Ravi Kumar Sharma[2,*]

[1]Department of Biochemistry, Faculty of Life Sciences, Aligarh Muslim University, Aligarh, India; [2]Botany Division, Central Drug Research Institute, M G Marg, Lucknow, India; Ravi Kumar Sharma* - E-mail: rkscdri@gmail.com; * Corresponding author

received December 03, 2007; revised December 08, 2007; accepted December 11, 2007; published online December 12, 2007

Abstract:
One of the most remarkable observations stemming from the sequencing of genomes of diverse species is that the number of protein-coding genes in an organism does not correlate with its overall cellular complexity. Alternative splicing, a key mechanism for generating protein complexity, has been suggested as one of the major explanation for this discrepancy between the number of genes and genome complexity. Determining the extent and importance of alternative splicing required the confluence of critical advances in data acquisition, improved understanding of biological processes and the development of fast and accurate computational analysis tools. Although many model organisms have now been completely sequenced, we are still very far from understanding the exact frequency of alternative splicing from these sequenced genomes.This paper will highlight some recent progress and future challenges for functional genomics and bioinformatics in this rapidly developing area.

Keywords: alternative splicing; protein; genes; cellular complexity

Background:

Alternative Splicing in various organisms

Alternative splicing is the process by which exons in transcripts are joined in different combinations to generate multiple mRNA variants, represents an important mechanism for the expression of structurally and functionally distinct proteins from a limited number of genes. **[1]** Recent genomic and bioinformatics analysis of vast amount of transcript data in human and other organisms suggest that alternative splicing is widespread in almost all higher eukaryotic genomes. **[2]** In humans, the frequency of alternative splicing has increased dramatically from a prior estimates as low as 5% **[3]** to 35-65% of all genes and 74% of all multi-exon genes having at least one alternative splice form. **[4]** Though, it is not known to what extent the resulting splice variants specify functionally relevant transcripts and proteins. This estimated prevalence of alternative splicing changes every year so 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.

Stetefeld and Ruegg, (2005) **[5]** estimated that almost 50% of eukaryotic genes and at least one-third of the genes if less complex organisms, such as nematode or flies, are alternatively spliced. It has been suggested that alternative splicing may not be as prevalent in plants as in mammals. Computational analysis of alternative splicing in *Arabidopsis thaliana* also questioned this assumption that alternative splicing is thought to be less prevalent in plants than it is in higher eukaryotes. **[6]** Moreover, recently, Ner-Gaon *et al.*, (2007) **[7]** showed that in comparison to higher animals, plants show a much greater degree of variety in their alternative splicing rates and in some plant species the rates of alternative splicing in animal and plant are very much comparable although the distribution of the types of spliced events may vary.

Brett *et al*., (2002) **[8]** in a large-scale expressed sequence tag (EST) analysis, estimated that, across a variety of distinct metazoan organisms as humans and nematodes, the rate of alternative splicing is similar. Kan *et al*., (2002) **[9]** have suggested that given a sufficient amount of EST coverage, alternative splice patterns may be observed for all genes that undergo splicing; hence, the higher the EST coverage, the higher the level of alternative splicing we expect. Contrary to this, Kim *et al*., (2004) **[10]** in a somewhat indirect method to calculate the level of alternative splicing, revealed that the level of alternative splicing varies between different organisms with a greater amount of alternative splicing in mammals compared with invertebrates. However, in reply, Harrington *et al*., (2004) **[11]** found that the most of the above studies were based on the EST coverage of the organisms, which has its own inherent limitations. Recently, Kim *et al*., (2007) **[12]** successfully demonstrated that vertebrates have a substantially higher

percentage of alternatively spliced genes compared with other species. Their results were based on a straight forward approach based on detection of alternative splicing events in gene-oriented clusters of mRNAs and ESTs in eight eukaryotic organisms**.** The lack of sufficiently large data sets of alternative splicing microarray data and sequenced ESTs and cDNAs has prevented reliable estimates of the proportions of genes that undergo alternative splicing in other organisms especially in *C. elegans*. Moreover, there are also insufficient data currently available to accurately assess the overall number of alternative splicing events in any one organism.

Major approaches used for identification of alternative spliced transcripts

Several different strategies have been applied to alternative splicing analysis, including (i) EST mapping against mRNA **[13]** (ii) mRNA/EST/protein mapping to the genome **[14]** (iii) splicing microarray analysis **[4]** (iv) *ab initio* machine learning approaches **[15,16]** and (v) Various other approaches published in recent years involved identifying and exploiting local sequence features for prediction. For instance, Dror *et al*., (2005) **[15]** used features like exon length, its divisibility by three, the length of flanking introns and the intensity of the polypyrimidine tract were utilized to identify possible spliced transcripts. Yeo *et al.*, (2005) **[17]** described an approach ACESCAN that is able to identify conserved exon skipping events in both human and mouse. This approach also uses exonic and intronic conservation as well as splice site scores, exon and intron lengths, and oligonucleotide composition. Ohler *et al*., (2005) **[18]** demonstrated that even alternative exons that are completely missed in current gene annotations can be discovered by applying a pair hidden Markov model algorithm to orthologous human-mouse introns. Ratsch *et al.*, (2005) **[16]** used a support vector machine to predict alternative exons. These studies demonstrate that a classifier based on characteristic genomic features can reliably predict exon skipping events ab initio. In-spite of a large pool of methods available, none of the approaches have been fully successful in delineating the full spectrum of alternative splice transcripts of a gene because of their inherent limitations or technical flaws. **[19, 20]**

Future directions:

The old axiom "one gene, one protein" no longer holds true, the more complex an organism, the more likely it became that way by extracting multiple protein meanings from individual genes. A number of bioinformatics analysis have been performed to elucidate the functional impact of alternative splicing but till now no single study has been fully successful in delineation of all possible spliced transcript of a gene. So, the major goal of alternative splicing annotation project in various organisms should be aimed at not only developing better and more efficient methods to identify the entire spectrum of alternative splice forms but also at developing an exhaustive pool of alternative transcripts of an organism to fully understand the complexity of eukaryotes.

References:

[01] D. L. Black, *Annu. Rev. Biochem.*, 72: 291 (2003) [PMID: 12626338]

[02] B. Modrek, *et al.*, *Nucleic Acids Res.*, 29: 2850 (2001) [PMID: 11433032]

[03] P. A. Sharp, *Cell*, 77: 805 (1994) [PMID: 7516265]

[04] J. Johnson, *et al.*, *Science*, 302: 2141 (2003) [PMID: 14684825]

[05] J. Stetefeld & M. A. Ruegg, *Trends Biochem. Sci.*, 30: 515 (2005) [PMID: 16023350]

[06] K. Iida, *et al.*, *Nucleic Acids Res.*, 32: 5096 (2004) [PMID: 15452276]

[07] H. Ner-Gaon, *et al.*, *Plant Physiology*, 144: 1632 (2007) [PMID: 17496110]

[08] D. Brett, *et al.*, *Nature Genet.*, 30: 29 (2002) [PMID: 11743582]

[09] Z. Kan, *et al.*, *Genome Res.*, 12: 1837 (2002) [PMID: 12466287]

[10] H. Kim, *et al.*, *Nat Genet.*, 36: 915 (2004) [PMID: 15340420]

[11] E. D. Harrington, *et al.*, *Nat Genet.*, 36: 915 (2004) [PMID: 15340420]

[12] E. Kim, *et al.*, *Nucleic Acids Res.*, 35: 125 (2007) [PMID: 17158149]

[13] D. Brett, *et al.*, *FEBS Lett.*, 474: 83 (2000) [PMID: 10828456]

[14] S. Gupta, *et al.*, *Bioinformatics*, 20: 2579 (2004) [PMID: 15117759]

[15] G. Dror, *et al.*, *Bioinformatics*, 21: 897 (2005) [PMID: 15531599]

[16] G. Ratsch, *et al.*, *Bioinformatics*, 21: 369 (2005) [PMID: 15961480]

[17] G. W. Yeo, *et al.*, *Proc. Natl Acad. Sci***.**, 102: 2850 (2005) [PMID: 15708978]

[18] U. Ohler, *et al.*, *PLoS Comp. Biol*., 1: 113 (2005) [PMID: 16110330]

[19] D. Talavera, *et al.*, *BMC Bioinformatics*, 8: 260 (2007) [PMID: 17640387]

[20] G. G. Leparc & R. D. Mitra, *Nucleic Acids Res.*, 35: 3192 (2007) [PMID: 17452356]

Edited by P. Kangueane

Citation: Kashyap & Sharma, *et al.*, Bioinformation 2(4): 155-156 (2007)