Alternative splicing and evolution

Stephanie Boue,* Ivica Letunic, and Peer Bork

Summary

Alternative splicing is a critical post-transcriptional event leading to an increase in the transcriptome diversity. Recent bioinformatics studies revealed a high frequency of alternative splicing. Although the extent of AS conservation among mammals is still being discussed, it has been argued that major forms of alternatively spliced transcripts are much better conserved than minor forms.[1] It suggests that alternative splicing plays a major role in genome evolution allowing new exons to evolve with less constraint. ***BioEssays*** **25:1031–1034, 2003. © 2003 Wiley Periodicals, Inc.**

Introduction

Anthropocentrism motivates human beings to search for an explanation of their relative complexity. This complexity used to be associated with the gene content of the genome. But the unexpectedly low number of genes revealed by the human genome sequencing project directed attention towards the mRNA and protein worlds.

Alternative splicing (AS) and alternative polyadenylation are two events that seem to increase protein diversity post-transcriptionally. AS mainly affects the coding region of the transcript whereas alternative polyadenylation modifies the 3′UTR, influencing the tissue distribution of the transcripts.[2] AS enables exons to be joined in different combinations and hence to produce distinct mature transcripts. AS has traditionally been thought of as an exceptional event occurring in only 5% of human genes,[3] but numerous computational biology studies in the last five years reveal a different picture.[4–12] In the absence of a reference dataset, it is difficult to compare the different approaches used, but the AS frequency estimated by different methods and various datasets is continuously increasing (Fig. 1). The most-common estimate nowadays indicates that 60% of all human genes have alternatively spliced variants, but a study of genes represented by more than 700 expressed sequence tags (ESTs) showed that 99% of them are subjected to AS.[10] Hence the estimation of 60% may increase as the size of EST databases grows.

Numerous case studies demonstrate the biological relevance of alternative splicing in health and in disease. AS increases transcriptome diversity and allows a specific expression of transcripts at precise time points in specific tissues. Its involvement in many diseases suggest that it would be a useful tool for disease diagnosis and treatment. In support of this is the use of modification of AS pathways as a chemotherapy approach by Mercatante and Kole.[13]

How did AS evolve to become a mechanism able to create complexity in higher organisms? We review here computational studies that used comparative genomics to propose some scenarios for the evolution of introns and AS.

First generation studies of alternative splicing: frequency estimation

Computational biologists possess different means to infer AS properties including two sources of expressed sequences. On the one hand, there are databases of mRNAs and full-length cDNAs whose quality is generally good but that are not numerous enough to reflect all the variety. On the other hand, there are databases of ESTs, relatively short sequences produced in a high throughput manner from many tissues, individuals and different conditions. This set represents an incomparable insight into the diversity produced in the cells, but might be biased towards medically relevant genes, is sometimes of poor quality, and may be contaminated by diverse vectors, non-processed mRNAs, or genomic sequences. Consequently, this great source of information must be handled with care and requires vigilant checking.

The first generation of large-scale AS studies was aimed either at estimating the frequency of the phenomenon in human or at discovering AS events happening in disease. ESTs or cDNAs are aligned to mRNAs or cDNAs, and gaps or insertions are sought to reveal AS events.[5,6,12] Different strategies are used to align expressed sequences to genomic sequences and retrieve AS events. An important point is that not only the degree of similarity of the sequences is taken into account, but also the presence of canonical splice sites at the

EMBL, Heidelberg, Germany.
*Correspondence to: Stephanie Boue, EMBL, Meyerhofstrasse, 69117 Heidelberg, Germany. E-mail: boue@embl.de
DOI 10.1002/bies.10371
Published online in Wiley InterScience (www.interscience.wiley.com).

Abbreviations: AS, alternative splicing; ESTs, expressed sequence tags; ESE, exon splicing enhancer; NMD, nonsense mediated decay; PTC, premature termination codon; SMART, Simple Modular Architecture Research Tool.

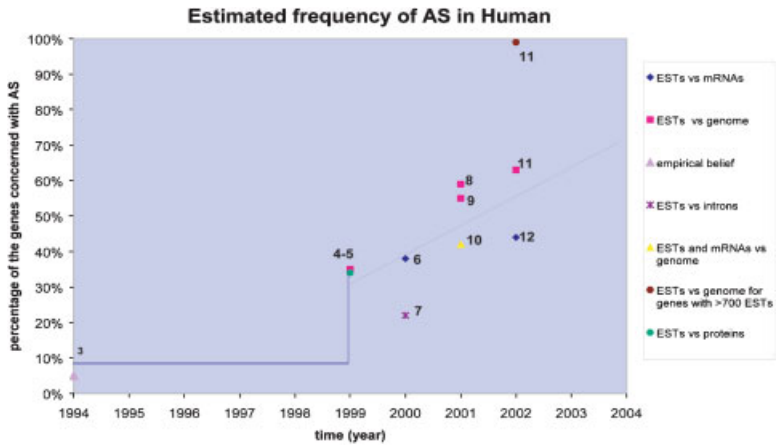

Figure 1. Evolution of the estimated frequency of AS over the last decade. Various computational biology methods (see legend) have been used to infer those numbers. The blue line shows the trend of the frequency evolution without taking into account the 99% found on a subset.

intron–exon borders. With the advent of whole genome sequencing projects, ESTs, cDNAs and mRNAs are aligned to genome sequences.[4,8–11,14] The high frequency of AS shown by those studies seems to explain complexity in higher organisms. To confirm this hypothesis, the frequency of AS was compared by Brett et al.[12] in seven species. ESTs were aligned to mRNAs for each species: the frequency of AS is similar in all of them. The observed differences of frequency can be attributed to a smaller EST coverage in some species. With genome sequences of several animals in hand, a second generation of computational studies aims at the conservation of AS in these species, and hence provides first insights into the evolution of AS.

Second generation studies: conservation of alternative splicing inferred by human/mouse comparison

A strategy to follow an AS event in two species is to look at AS patterns in orthologous genes in both species. Orthologs are homologous genes that evolved from a single ancestral gene in the last common ancestor of compared genomes. Orthology information can be retrieved from specialized databases (e.g., Homologene, http://www.ncbi.nlm.nih.gov/HomoloGene/).

A comparative analysis of 117 human and mouse orthologous gene pairs[15] shows that 95% of the genes have the same number of exons. Moreover, the length of corresponding exons is strongly conserved (identical in 73% of the cases) whereas the length of introns varies considerably (on average human introns are 1.5 times longer than mouse ones). Thus the differences between human and mouse transcripts are to be found either in the 5% of genes in which the number of exons is not conserved, or in differential splicing patterns. Actually, in 5% of the cases, one mouse exon corresponds to two human exons produced by exon duplication. Interestingly exon duplication is proven by other independent analyses[16,17] to be concomitant to alternative splicing. Thus, it is presumably a combination of gene structure and AS patterns changes that explains the gain of complexity.

To infer the role of alternative splicing in evolution, one has to distinguish between alternative and constitutive exons. Constitutive exons are found in every transcript of a gene whereas alternative exons are absent in at least one transcript that would have been long enough to contain it. Modrek and Lee[1] analyze a set of 9,434 orthologous genes in human and mouse. Within each gene pair, they assign orthologous exons whenever possible: 90% of the exons are assigned to an orthologous pair. Those exons show a high level of similarity (87% of identity). This work introduces an additional distinction within alternative exons: "major form" if the exon appears in at least 50% of the transcripts and "minor form" otherwise. The track of all exons reveals high conservation of the constitutive exons (98%) as well as the major forms of alternative exons (98%). Interestingly, minor forms of alternative exons are much less conserved (only 25% of conservation). Although the study was conducted carefully enough to discard possible contaminations, an important functional question is whether these minor forms are real exons rather than mispredictions. Actually the probability of rare spliceosomal error is low because the majority of those minor exons is supported by many ESTs. This work leads to the conclusions that orthologous exons are similarly regulated and that the inclusion level of an exon in human ESTs accurately predicts the inclusion level of the orthologous exon in mouse. In other words, if an exon is constitutive (100% inclusion) or the major form of an alternative exon ($\geq$50% inclusion) the probability is very high that it will be included in most of the mouse transcripts as well. To validate the strategy used for the mouse/human comparison, it is applied to rat/human comparison as well. The rat exons are identified using rat ESTs and mRNAs and used to look into the human genome for orthologous exons. The results are similar

to the previous ones, revealing 86–90% conservation for major forms and only 16–33% for minor.

A concurrent study of 166 pairs of orthologous alternatively spliced genes[18] claims to find a low conservation of AS patterns between human and mouse. As databases are not exhaustive, Nurtdinov et al. limit their analysis to known cassette exons, retained introns and alternative splice sites and their conservation in the genomes. According to the authors, a conservative estimation shows that approximately 50% of alternatively spliced genes have non-conserved isoforms

A third investigation by Thanaraj et al.[19] based on exon junctions estimates an overall conservation of 61% of alternative and 74% of constitutive splice junctions. Those numbers are once again dependent on the coverage of those junctions by expressed sequences and are considered as minima.

At a first glance, the numbers revealed by these studies vary considerably. However, the differences observed can be explained partly by the use of different datasets and partly by the degree of conservatism of the work. As a conclusion, many AS events are conserved, with the exception of minor forms, which moreover are often associated with exon creation or loss. Such an exon loss event in a human gene, which correlates with an AS event in mouse (Lactadherin) is illustrated in Fig. 2.

Large-scale experimental validation is still lacking to definitely confirm these predictions. New technologies such as microarrays to determine gene structure are now appearing,[20] suggesting that this validation step will be available soon. Until then, as species-specific isoforms are generally thought to be expressed tissue-specifically and at low levels, it is most probable that we miss many variants. However, many of the predicted or observed isoforms are likely to be non-functional as they are truncated or deleted within functional domains. Thus the numbers found in all those studies are to be regarded with caution and as non definite.

Towards ab initio prediction of alternative splicing?

In order to predict alternative splice variants ab initio rather than to detect them by comparative analyses, some groups look for signals contained in DNA sequences that are recognized by the cell to differentiate constitutive exons from alternative ones. Splice sites and other consensus sequences involved in splicing are compared between human and mouse.[22] It has already been shown that alternative exons possess weaker splice sites.[23] To go further, Modrek and Lee distinguished major and minor forms of alternative exons. However, no differential signals are found around the splice sites. The distinction may thus be present within exonic sequence enhancers (ESEs). Prediction of AS events is not yet possible. We still need a better comprehension of the splicing mechanism itself and its regulation to achieve it.

AS and evolution

The birth of introns is one of the most intriguing phenomena in genome evolution. Lynch and Kewalramani[24] investigate the proliferation of introns and link the apparition and multiplication of introns in genomes with nonsense mediated decay (NMD). NMD directs mRNAs containing a premature termination codon (PTC) to degradation, and hence prevents the cell from producing non-functional proteins. A priori, newly arisen introns are deleterious for the gene and therefore should be subjected to selective pressure. But, by decreasing the selective pressure on introns, NMD acts as a facilitator of intron

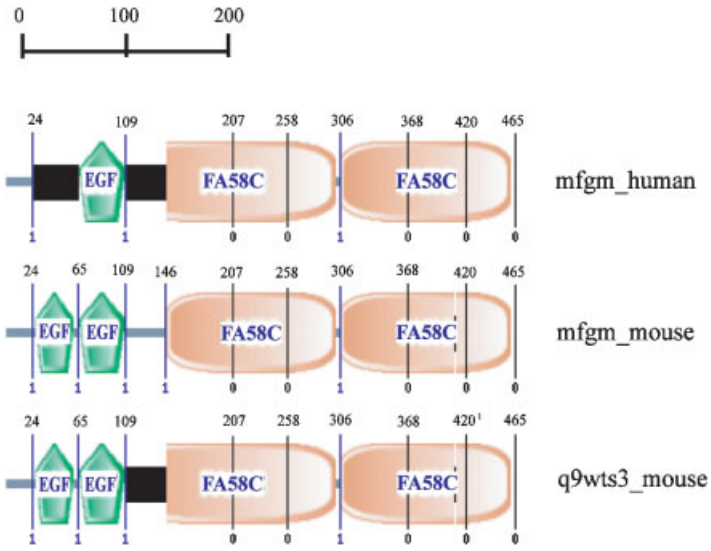

Figure 2. Example of exon loss coupled with AS: Lactadherin. Lactadherin belongs to a family of mosaic glycoproteins widely expressed in a number of tissues and body fluids. The figure shows SMART[21] protein representations with included intron positions (vertical lines) mapped onto a ClustalW alignment of the orthologous genes in human and mouse (mfgm_human and mfgm_mouse). An AS variant in mouse (q9wts3_mouse) is shown to skip the exon which is missing in human. This event is proven to be an exon loss in human rather than a gain in mouse because this exon is found in other organisms. Black boxes represent gaps in the alignment.

proliferation. However, the number of introns per gene cannot be infinite. So the rate of colonization of introns is expected to be negatively associated with intron number: once enough coverage for NMD has been acquired, all newly introns are weakly selected against.

Lewis et al.[25] observe that alternative splicing and NMD are widely coupled. One third of AS variants that they examined contained PTC, making them targets for NMD. This phenomenon should not be seen as a cellular waste because unusable transcripts are produced, but rather as a means for the cell to regulate gene expression at the post-transcriptional level in a tissue- or time-specific manner.

Modrek and Lee[1] propose a concept of AS evolution based on the fitness landscape and adaptive walks theory. According to this theory, the fitness of an organism is determined by two factors: the internal state of the organism and the environment in which it lives. The organisms are defined by any two characteristics and placed in a two-dimensional plane according to them. Fitness is introduced by turning the surface from a plane into a more rugged landscape with the peaks corresponding to high fitness and the valleys to low fitness. The only way that a population can change its location on the landscape is to have offspring with different genotypes to their parents. However, each step on the landscape has to be uphill in the direction of higher fitness or it would produce organisms less well adapted to their environment with less chance of survival. That is where AS comes into play. If a new exon is incorporated into a gene and alternatively spliced, it would probably first be included in only few of the transcripts and would be free to evolve as the original transcript form would still accomplish its function. In this sense, alternative splicing allows an organism to convert forms with low fitness (inclusion of a new exon) to higher fitness (after accumulation of mutations creating a new useful function). Without AS, those new exons would probably be selected against, lowering the capacity for evolution of an organism. This hypothesis is supported by various types of evidence. Indeed two phenomena are already known to produce new exons that are often alternatively spliced: exon duplication[16,17] and alternative 3′ splice site selection within Alu elements.[26]

Conclusion

Recent computational predictions have revealed the high frequency and biological relevance of AS. Some current bioinformatics analyses are now dedicated to discover the benefits of AS for an organism in evolutionary terms, which lead to its spreading. AS thus plays at least two roles. On the one hand, it is a mechanism able to produce in an economical way diversity and specificity at the cell-, tissue- or developmental levels. On the other hand, by decreasing, together with NMD, the selective pressure on genes, it seems to allow a trial/error approach for the evolution of the gene structure.

References

1. Modrek B, Lee CJ. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet 2003;34:177–180.
2. Beaudoing E, Gautheret D. Identification of alternative polyadenylation sites and analysis of their tissue distribution using EST data. Genome Res 2001;11:1520–1526.
3. Sharp PA. Split genes and RNA splicing. Cell 1994;77:805–815.
4. Mironov AA, Fickett JW, Gelfand MS. Frequent alternative splicing of human genes. Genome Res 1999;9:1288–1293.
5. Hanke J, Brett D, Zastrow I, Aydin A, Delbrück S, Lehmann G, Luft F, Reich J, Bork P. Alternative splicing of human genes—more the rule than the exception? Trends Genet 1999;15:389–390.
6. Brett D, Hanke J, Lehmann G, Haase S, Delbrück S, Krueger S, Reich J, Bork P. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett 2000;474:83–86.
7. Croft L, Schandorff S, Clark F, Burrage K, Arctander P, Mattick JS. ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. Nat Genet 2000;24:340–341.
8. Consortium IHG. Initial sequencing and analysis of the human genome. Nature 2001;409:860–924.
9. Kan Z, Rouchka EC, Gish WR, States DJ. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res 2001;11:889–900.
10. Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res 2001:2850–2859.
11. Kan Z, States DJ, Gish WR. Selecting for functional alternative splices in ESTs. Genome Res 2002;12:1837–1845.
12. Brett D, Pospisil H, Valcartel J, Reich J, Bork P. Alternative splicing and genome complexity. Nat Genet 2002;30:29–30.
13. Mercatante DR, Kole R. Control of alternative splicing by antisense oligonucleotides as a potential chemotherapy: effects on gene expression. Biochim Biophys Acta 2002;1587:126–132.
14. Zavolan M, Kondo S, Schönbach C, Adachi J, Hume DA, Group RG, members G, Hayashizaki Y, Gaasterland T. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res 2003;13:1290–1300.
15. Batzoglou S, Patcher L, Mesirov JP, Berger B, Lander ES. Human and mouse gene structure: comparative analysis and application to exon prediction. Genome Res 2000;10:950–958.
16. Kondrashov FA, Koonin EV. Origin of alternative splicing by tandem exon duplication. Hum Mol Genet 2001;10:2661–2669.
17. Letunic I, Copley RR, Bork P. Common exon duplication in animals and its role in alternative splicing. Hum Mol Genet 2002;11:1561–1567.
18. Nurtdinov RN, Artamonova II, Mironov AA, Gelfand MS. Low conservation of alternative splicing patterns in the human and mouse genomes. Hum Mol Genet 2003;12:1313–1320.
19. Thanaraj TA, Clark F, Muilu J. Conservation of human alternative splice events in mouse. Nucleic Acids Res 2003;31:2544–2552.
20. Wang H, et al. Gene structure-based splice variant deconvolution using a microarray platform. Bioinformatics 2003;19:i315–i322.
21. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P. Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res 2002;30:242–244.
22. Sorek R, Ast G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res 2003;13:1631–1637.
23. Stamm S, Zhang MQ, Marr TG, Helfman DM. A sequence compilation and comparison of exons that are alternatively spliced in neurons. Nucleic Acids Res 1994;9:1515–1526.
24. Lynch M, Kewalramani A. Messenger RNA surveillance and the evolutionary proliferation of introns. Mol Biol Evol 2003;20:563–571.
25. Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay. Proc Natl Acad Sci USA 2003;100:189–192.
26. Lev-Maor G, Sorek R, Shomron N, Ast G. The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. Science 2003;300:1288–1291.