Alternative splicing and genome complexity

Published online: 17 December 2001, DOI: 10.1038/ng803

Alternative splicing of mRNA allows many gene products with different functions to be produced from a single coding sequence. It has recently been proposed as a mechanism by which higher-order diversity is generated. Here we show, using large-scale expressed sequence tag (EST) analysis, that among seven different eukaryotes the amount of alternative splicing is comparable, with no large differences between humans and other animals.

The unexpectedly low number of genes identified in the human genome raises again the question of the source of an organism's complexity[1,2]. One possible source is the greater complexity of the human genes (that is, their modularity) compared with those of other multicellular organisms; this observation suggests a much higher level of regulation of genes and pathways[1,2]. Another source may be post-translational modifications; more than 200 different types are known, and it is predicted that, on average, for each human gene three different modified proteins with different functions are produced (reviewed in ref. 3). Finally, alternative splicing of human genes might provide many more proteins per gene than in other organisms[4,6]. Although it has long been assumed that only five percent of human genes are alternatively spliced[7], more recent estimates based on ESTs mapped onto mRNA sequences indicate a much higher rate of alternative splicing in human genes[8–10] . It has also become clear that the estimates increase with EST coverage (Fig. 1)[6,11], and, by including human genomic data, it was predicted that at least 50% of the human genes are subjected to alternative splicing[1,2]. Here we show that this fraction is likely to be similar in other animals, including invertebrates, and suggest that the relatively low number of human genes identified is not accompanied by a uniquely high rate of alternative splicing. We compared seven eukaryotic organisms, with sufficient coverage of ESTs and mRNA data, using a bioinformatics protocol that detects possible alternative splice forms by comparing high-scoring ESTs to mRNA sequences using BLAST[9]. We have written filtering programs that compare the ends of each aligned sequence pair for deletions or insertions in the EST sequence, which suggest the existence of alternative splice forms. The method has been validated previously by RT–PCR and subsequent sequencing[9,11,12] and has already been used to successfully predict colon cancer–specific alternative splicing[12]. All together[9,11,12], 35 predicted alternative-splicing candidates were tested with a success rate of 92% (ref. 11). Estimates of 2.5% of false negatives due to DNA contamination of mRNA were derived from a sample of 120 genes[9].

In light of the differences in mRNA and EST coverage among species, we created subsets of the data derived from EST comparisons for each organism (Fig. 1). When we used these comparison sets (and other subsets generated for statistical support), all vertebrates and invertebrates showed a similar rate of alternative splicing with respect to both the number of genes affected and the number of variants per gene. Only *Arabidopsis thaliana* showed a lower rate (Fig. 1). Evidently, the use of ESTs to estimate rates of alternative splicing is limited by, for example, overrepresentation of 3′ untranslated regions, biases towards widely expressed genes and the likelihood of some false-negative and false-positive predictions. Although a much greater number and diversity of ESTs is required to detect the vast majority of splice sites in each species (current estimates have to be seen as lower limits; Fig. 2), the similarity of the rates in the different species should be only marginally affected. The small dif-

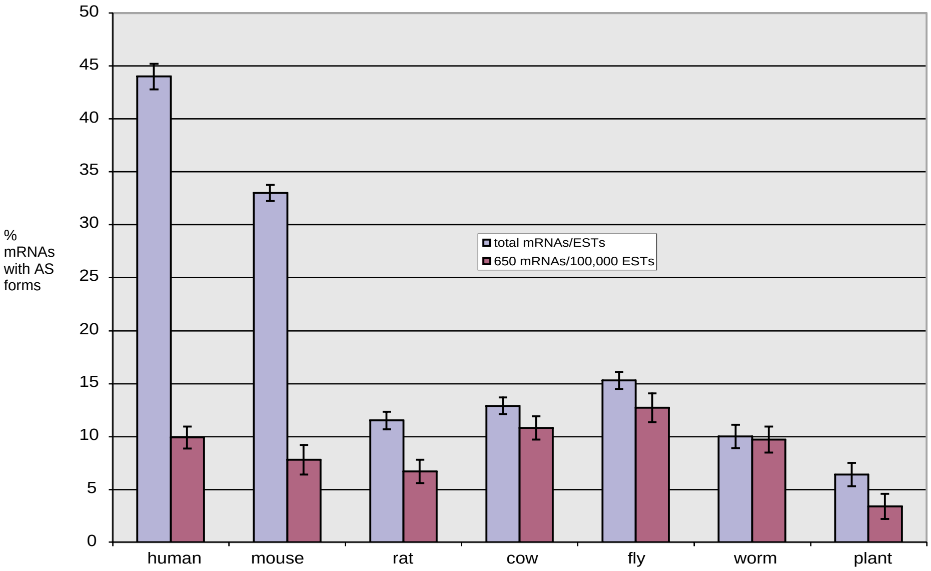
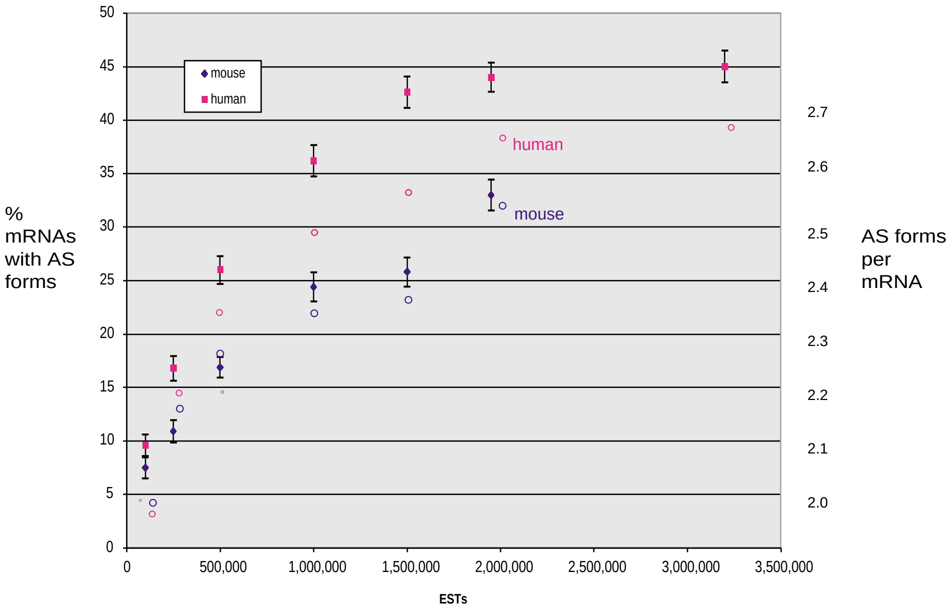


Fig. 1 EST estimations of alternative splicing from different eurkaryotes. To identify alternative splicing (AS), we identified high-scoring ESTs with more than 98% identity over 100 bps to mRNA sequences (mRNAs), using the BLASTN search tool with slightly modified parameters[8,11]. We recorded deleted or inserted sequences in each matching EST after filtering internal repeats. The bar to the left (light purple) for each organism shows the estimate of alternative splicing based on all ESTs and the total number of published mRNA sequences and ESTs for the species: *Homo sapiens*, 23,161 mRNAs and 3.1 million ESTs; *Mus musculus*, 9,682 mRNAs and 1.9 million ESTs; *Rattus norvegicus*, 5,803 mRNAs and 263,362 ESTs; *Drosophila melanogaster*, 2,973 mRNAs and 115,191 ESTs; *Bos taurus*, 1,370 mRNAs and 159,130 ESTs; *Caenorhabditis elegans*, 18,821 mRNAs and 108,115 ESTs; *Arabidopsis thaliana*, 3,084 mRNAs and 112,112 ESTs. As the rate of detectable alternative splicing depends on EST coverage, we created comparable subsets for each organism, comprising a random set of 650 mRNA sequences with a coverage of 100,000 ESTs (to allow inclusion of cow with the smallest number of public mRNA sequences and ESTs of the seven organisms). Only mRNAs with 3–20 EST matches were included in the set, to account for biases in EST coverage. The darker red bar to the right for each organism is the outcome of this subset. Error bars were created with normal binomial s.d. We carried out random re-sampling with smaller subsets of ESTs (100,000) tested against the total mRNA sets. We observed a similar distribution and scale of error bars.

Fig. 2 Dependence of alternative splicing prediction on EST coverage. We calculated the percentages of alternatively spliced (AS) mRNA sequences (mRNAs) for human and mouse, using the set of 650 mRNA sequences from Fig. 1 and increasing the number of randomly selected ESTs to the currently available maximum for both species. Error bars were created as in Fig. 1. The open circles (pink or blue) on the graph represent the number of different alternative splice forms per mRNA found from matching ESTs in mouse and human (right axis). Note that the saturation indicated in the figure remains a lower estimate because of factors[9,11] such as limited EST coverage per position and limited cDNA library variability per gene.



ferences observed across the animals probably result from differences in library coverage in the different EST data sets. For some species, such as human and mouse, a much wider selection of tissues is available in the EST database than, for example, fruitfly or cow. Although elaborate combinations of patterns of alternative splicing in specific genes (such as cell adhesion molecules, ion channels or proteins that determine cell shape; all have several alternatively spliced forms) can contribute to the complexity of a system or organism, the majority of genes in the animals studied here have similar numbers of alternatively spliced forms per gene. We therefore conclude that although alternative splicing can significantly expand the coding capacity of genomes, similar levels of alternative splicing across species argues against an overall increase in splicing as a source of increase in genome and organism complexity. The data also suggest that there are a wide variety of novel gene products to be investigated in all animals that are further diversified by post-translational modification and other processes.

URLs. The splice-site database containing more than 12,500 potential alternative splicing sites from seven organisms, after filtering redundant mRNA sequences including all the sites discussed here: http://rhodos.bioinf.mdc-berlin.de/asforms.

David Brett[1], Heike Pospisil[1], Juan Valcárcel[2], Jens Reich[1] & Peer Bork[1,2]

[1]*Max Delbrück Center for Molecular Medicine, Robert-Rössle Strasse 10, Berlin-Buch, 13125 Germany.* [2]*European Molecular Biology Laboratory, Meyerhofstr. 1, 69012 Heidelberg, Germany.*

Received 18 July; accepted 26 November 2001.

1. The International Genome Sequencing Consortium. *Nature* **409**, 860–921 (2001).
2. Venter, C. *et al. Science* **16**,1304–1351 (2001).
3. Banks, RE. *et al. Lancet* **18**, 1749–1756 (2000).
4. Ewing, B. & Green, P. *Nature Genet.* **25**, 232–234 (2000).
5. Crollius, H.R. *et al. Nature Genet.* **25**, 235–238 (2000).
6. Kan, Z. *et al. Genome Res.* **11**, 889–900 (2001).
7. Sharp, P.A. *Cell* **77**, 805–815 (1994).
8. Mironov, A.A. *et al. Genome Res.* **9**, 1288–1293 (1999).
9. Hanke, J. *et al. Trends Genet.* **15**, 389–390 (1999).
10. Croft, L. *et al. Nature Genet.* **24**, 340–341 (2000).
11. Brett, D. *et al. FEBS Lett.* **26**, 83–86 (2000).
12. Brett, D. *et al. Oncogene* **20**, 4581–4585 (2001).