Positive Selection and Alternative Splicing in Humans

Mikhail S Gelfand, *Institute for Information Transmission Problems (the Kharkevich Institute), RAS, Moscow, Russia*

Vasily E Ramensky, *Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia*

Advanced article

Article Contents

Online posting date: 15th December 2009

Alternative splicing is an important mechanism of generating protein diversity and accelerated genome evolution. The mode of the selection acting in constitutive, major alternative and minor alternative regions of human genes is different. Whereas constitutive and major alternative regions tend to evolve under negative (stabilizing) selection, alternatively spliced exons from minor isoforms experience lower selective pressure at the amino acid level accompanied by weak selection against synonymous sequence variation. The McDonald–Kreitman test uses the nucleotide variation for a gene or a set of genes between and within species to detect the positive Darwinian selection in the presence of negative selection. The results of the test suggest that alternatively spliced exons are also subject to positive selection, with up to 27% of amino acids fixed by positive selection.

Functional and Evolutionary Roles of Alternative Splicing

Alternative splicing (AS) is one of the major mechanisms for generating protein diversity in eukaryotes (Graveley, 2001). By recent estimates, it affects approximately 90% of the human genes (Wang *et al*., 2008), with the prevalence of alternative splicing in invertebrates and plants being lower (Ner-Gaon *et al*., 2007; Kim *et al*., 2004, 2007; Brett *et al*., 2002), and in protists, almost negligible. Alternative splicing is also important for maintaining protein identity: if a cell needs two proteins that are exactly identical in some domains and differ at others (e.g. membrane-bound antigen receptors and secreted immunoglobulins in B-lymphocytes), these isoforms are created by alternative splicing rather than gene duplication.

ELS subject area: Evolution and Diversity of Life

How to cite:
Gelfand, Mikhail S; and Ramensky, Vasily E (December 2009) Positive Selection and Alternative Splicing in Humans. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0021735

In addition to its role in development, stage- and tissue-specific regulation of gene expression, alternative splicing is important for generating protein diversity in evolution. Indeed, it has been repeatedly demonstrated that alternatively spliced exons and alternative splice sites are less conserved than constitutive ones in the human–mouse comparisons (Nurtdinov *et al*., 2003; Modrek and Lee, 2003) and in the comparison of two *Drosophila* genomes and the malarial mosquito (Malko *et al*., 2006), although not in nematodes (Irimia *et al*., 2008). Moreover, multiple comparisons allowed for distinction between gain and loss of exons and demonstrated that alternatively spliced exons indeed are often young rather than simply deteriorating (Nurtdinov *et al*., 2007).

An important arising question was whether species-specific exons are functional or simply represent splicing noise. The latter theory was supported by the observation that many species-specific alternative exons disrupt the reading frame (Sorek *et al*., 2004). This is an open question, since one possibility is that these isoforms represent regulatory sinks: splice products specifically targeted for nonsense-mediated decay (McGlincy and Smith, 2008; Severing *et al*., 2009).

The raw material for these young introns is provided by cryptic splice sites and repeats. Indeed, approximately 50% of disease-causing mutations disrupting splice sites result in activation of cryptic sites, both donor (approximately 40%) and acceptor (approximately 60%) (Kurmangaliyev and Gelfand, 2008); mutations may also activate cryptic sites by improving their fit to the consensus (Královicová and Vorechovsky, 2007). One can follow the birth of splice sites creating new, alternatively spliced exons by comparing the recently duplicated genes. One such gene family is human MAGE-A (melanoma-associated antigen A; Artamonova and Gelfand, 2004). The founder gene of this paralogous family was a retrogene that had no introns. After several duplication events, the descendant genes accumulated point differences, some of which created splice sites in the noncoding 5′ region. When a suitable pre-existing cryptic site was available, this yielded a new, alternative cassette exon.

A rich source of cryptic splice sites are repeats, in particular, Alu repeats. Indeed, dependent on the stringency of search criteria, 0.1–4% human mature messenger ribonucleic acids (mRNAs) were shown to contain Alu repeats in the protein-coding region (Gotea and Makałowski, 2006). The Alu consensus contains several cryptic sites, both donor and acceptor. Activation of such sites, e.g. by point mutations, creates exon extensions, and in all known cases such sites are alternative (Lev-Maor *et al.*, 2003; Sorek *et al.*, 2002). If cryptic sites of both types are utilized, integration of Alu may create a cassette exon (Zhang and Chasin, 2006). SINE (short interspersed nuclear element) retrotransposons B4 and MIR (mammalian interspersed repeat) also have been implicated in alternative splicing (Zheng *et al.*, 2005).

One more source of new exons is exon duplication. Duplicated exons were observed in 11% of human genes and 7–8% of nematode (*Caenorhabditis elegans*) and fruitfly (*Drosophila melanogaster*) genes. The length of approximately 60% of duplicated exons is not divisible by 3 (Letunic *et al.*, 2002) and such exons are likely alternatively spliced, since their simultaneous incorporation into mRNA would disrupt the reading frame downstream.

Hence, many, if not most, young exons are alternative. However, one would expect new exons or exon extensions to evolve under positive selection. Indeed, if a protein-coding fragment has been formed from a previously noncoding region, and has become fixed due to some advantage it provided, it is very likely that this fragment is suboptimal, and it would accumulate mutations to improve its functionality. However, recently duplicated exons may be fixed if they accumulate differences, similarly to recently duplicated genes (Jordan *et al.*, 2004). One example is provided by ion channel genes where incorporation of mutually exclusive exons yields proteins with different properties (Stamm *et al.*, 2005). **See also**: Alternative Splicing in the Human Genome and its Evolutionary Consequences; Alternative Splicing: Cell-type-specific and Developmental Control; Alternative Splicing: Evolution

Kn/*Ks* Test

The way to identify positive selection in protein-coding regions is to compare the nonsynonymous with synonymous substitution rates (*K*n and *K*s, respectively). In a neutrally evolving region these rates, if correctly normalized for the number of synonymous and nonsynonymous sites, respectively, are equal. Positive selection manifests as *K*n > *K*s, whereas in negatively selected regions *K*n < *K*s.

However, there are several problems with this approach. Firstly, it is insufficiently sensitive: only rarely one observes so strong and ubiquitous positive selection, that it yields *K*n > *K*s. Indeed, if only a relatively small fraction of the sites evolves under positive selection, whereas at the majority of the sites nonsynonymous substitutions are mildly deleterious, they would dominate the rates, and one would still observe *K*n < *K*s. One way to avoid this problem is to compute the *K*n/*K*s ratios at individual sites, but this requires multiple alignments containing many sequences of orthologous genes from sufficiently close species (otherwise, when multiple substitutions have occurred at one site, the *K*n value becomes unreliable).

One may simply compare the average *K*n and *K*s values and their ratio in two groups of sequences, for example, in constitutive and alternative regions. Indeed, it has been observed that alternative splicing relaxes the selection pressure against amino acid substitutions, leading to higher *K*n in alternative regions compared to constitutive ones (Artamonova and Gelfand, 2004; Ermakova *et al.*, 2006). However, these observations may not be taken as firm evidence of positive selection, since the same would be expected under relaxation of negative selection. Moreover, the *K*n/*K*s test fails for one additional reason: somewhat unexpectedly, *K*s is generally higher in alternative regions than in constitutive ones and thus is nonneutral. The possible reason for that is that alternative exons contain more splicing regulatory sites, enhancers and silencers, than constitutive ones. Indeed, the *K*s rates have been shown to be lower near splice sites (Baek and Green, 2005; Parmley and Hurst, 2007; Xing *et al.*, 2006), in exons flanking long introns (Dewey *et al.*, 2006) and in regions involved in complex alternative splicing as compared to simple cassette exons and exon extensions. There are also additional confounding effects: the evolutionary rates are different for the major and minor isoform cassette exons (Xing and Lee, 2005), for *N*-terminal, internal and *C*-terminal alternative regions (Ermakova *et al.*, 2006) and for different types of elementary alternatives (Malko *et al.*, 2009). **See also**: Synonymous and Nonsynonymous Rates

McDonald–Kreitman Test

Although there exist several recently developed approaches aiming at identification of positive selection signatures in the human genome (Sabeti *et al.*, 2006, 2007), they are not applicable to the case of closely intermixed genomic regions, characteristic of genomic AS. However, the availability of the SNP (single nucleotide protein) data makes it possible to apply more advanced analysis via the McDonald–Kreitman test (McDonald and Kreitman, 1991), which enables the detection of positive Darwinian selection in the presence of negative selection (Fay *et al.*, 2002; Bustamante *et al.*, 2005). This test relies on counting inter- and intra-species nucleotide differences for a gene or a set of genes between and within species. More precisely, the numbers of synonymous polymorphisms (Ps) and substitutions (Ds) and the numbers of nonsynonymous polymorphisms (Pa) and substitutions (Da) are compared in a contingency table. A statistically significant excess of nonsynonymous substitutions relative to polymorphisms (Da/Ds > Pa/Ps) implies positive selection that provides fixation of advantageous mutations. The fraction α of fixed amino acid substitutions that cannot be explained by neutral variation and hence are presumably driven by positive selection is then estimated as $\alpha = 1 - (\mathrm{Pa/Ps})/(\mathrm{Da/Ds})$ (Smith and Eyre-Walker, 2002).

Later, we describe the differences between the rates of synonymous and nonsynonymous polymorphism and human–chimpanzee divergence observed in alternative and constitutive regions of 6672 human genes (Ramensky *et al.*, 2008). These genes contain 52 151 constitutive and 14 196 alternative exons. Depending on the total number of proteins, mRNA and EST sequences that cover the corresponding exon the latter could be classified into major (included into at least 2/3 of all isoforms) and minor (included into less than 2/3 of all isoforms) ones. Exons were defined as conserved if they had donor and acceptor sites conserved in at least one of the orthologous mouse and dog genes, and nonconserved if either the exon or the entire genes were not conserved in both dog and mouse. The genes harbour 6465 SNPs and 50 649 human–chimpanzee substitutions derived from the whole genome alignments.

The ratios *K*n/*K*s for SNPs and divergence (**Table 1**) are very close for constitutive and major alternative regions but differ more than 2-fold from those for functional minor alternative regions, confirming that lower negative selection against amino acid substitutions is characteristic of these fragments. The fraction α of fixed amino acid substitutions driven by positive selection equals 0.27 for the minor alternatives suggesting that they experience positive selection, unlike the major alternatives and constitutive exons with $\alpha < 0$ indicating purifying selection (**Table 1a**). The Fisher's test value, *F*, given in the last column of the table reflects the probability that the difference between Da/Ds and Pa/Ps is random. Positive selection in the minor alternatives is still observed when the exons are split into conserved ($\alpha = 0.35$, **Table 1b**) and nonconserved ($\alpha = 0.23$, **Table 1c**).

This is consistent with a notion that new exons emerge by fixation of aberrant splicing events with subsequent upregulation of functionally useful variants. Indeed, since minor alternative exons, unlike constitutive and major alternative ones, seem to be relatively young (Zhang and Chasin, 2006), they are a natural substrate for positive selection. The α values in the constitutive and major alternative exons differ between these two groups, with negative selection ($\alpha < 0$) characteristic for the conserved exons and positive selection ($\alpha > 0$) dominating in the nonconserved ones. Thus, all types of nonconserved exons seem to experience positive selection. Unexpectedly, the fraction of amino acid substitutions fixed by positive selection in the conserved minor alternative exons is higher than that in nonconserved exons ($\alpha = 0.35$ versus $\alpha = 0.23$). This fact may be explained by contamination of the minor isoform sample by aberrantly spliced, nonfunctional exons, blurring the evidence for selection. Indeed,

Table 1 SNP and divergence density values and the McDonald–Kreitman test

Exon type	Number of exons	*K*n/*K*s (SNP)	*K*n/*K*s (Div)	α	The Fisher's test significance
(a) All exons					
Constitutive	52 151	0.26	0.25	−0.04	0.118
Major	9799	0.26	0.26	−0.02	0.442
Minor	4397	0.55	0.76	0.27	0.001
(b) Conserved exons					
Constitutive	42 590	0.25	0.23	−0.09	0.008
Major	7218	0.26	0.23	−0.16	0.069
Minor	2023	0.40	0.62	0.35	0.009
(c) Nonconserved exons					
Constitutive	9561	0.29	0.32	0.09	0.056
Major	2581	0.27	0.34	0.21	0.052
Minor	2374	0.66	0.85	0.23	0.026
(d) All exons, SNPs with derived allele frequency $\geq 5\%$					
Constitutive	30 060	0.25	0.25	0.04	0.149
Major	4801	0.23	0.29	0.19	0.015
Minor	2026	0.54	0.74	0.27	0.012

Notes: The table contains the ratio values *K*n/*K*s of nonsynonymous substitutions per nonsynonymous site (*K*n) to synonymous substitutions per synonymous site (*K*s) for three types of gene regions. *K*n and *K*s denote both polymorphism densities and divergence, with discriminating labels 'SNP' and 'Div', respectively. The last column shows the significance computed by the Fisher's test applied to the 4 numbers organized in a 2-by-2 contingency table. (a) All exons; (b) Conserved exons; (c) Nonconserved exons from genes covered by ≥ 60 ESTs and (d) Genes with at least one SNP with known frequency, only neutral SNPs (both validated and nonvalidated) with the frequency of the derived (new) allele $\geq 5\%$.
Source: Reproduced from Ramensky *et al.* (2008). With permission from Elsevier.

nonconserved minor alternative exons are always suspicious (Sorek *et al.*, 2004), and thus this result should be re-examined as more data become available.

The McDonald–Kreitman test relies on the assumption that the polymorphism accounted for is neutral. To eliminate potentially mildly deleterious SNPs that might have accumulated in the human population, for example, as a result of rapid population expansion, the test was performed with the subset of SNPs for which the frequency of the derived (new) allele is known and not less than 5% (**Table 1d**). With these presumably neutral SNPs, the fraction α of substitutions fixed by positive selection for minor alternatives is estimated as 0.27, with lower statistical reliability (Fisher's test value approximately 0.01). Positive selection ($\alpha = 0.19$) also appears in the major regions suggesting that in the total SNP sample weaker positive selection in the major exons is masked by the presence of nonneutral SNPs.

Positive selection could not be observed in earlier studies relying on comparisons of Kn/Ks values in the alternative and constitutive regions of pair-wise aligned mammalian genes (Ermakova *et al.*, 2006), since the latter technique does not allow one to distinguish between positive selection and relaxed negative selection. Here, the increase of nonsynonymous diversity in minor isoforms is greater than that expected due to the relaxation of negative selection estimated from the polymorphism levels. Hence this increase can naturally be interpreted as a trace of positive selection.

The results outlined here have been validated on different sets of SNPs (Ramensky *et al.*, 2008). One should be aware, however, that the nonsystematic nature of the polymorphism data may be a potential source of artifactual effects. Besides, as more mammalian genomes are being sequenced, an opportunity to use alternative methods for estimating positive selection based on representative multiple alignments would emerge. **See also**: Gene Evolution and Human Adaptation; Identifying Regions of the Human Genome that Exhibit Evidence for Positive Selection; Neutrality and Selection in Molecular Evolution: Statistical Tests; Purifying Selection: Action on Silent Sites; Selection Operating on Protein-coding Genes in the Human Genome

Conclusion

Positive selection shown to operate at about a quarter of positions in the minor alternative regions is consistent with the theory that many such exons are relatively young, and hence not optimal. However, the constitutive and major alternative regions are similar and evolve mainly under negative selection. Again, it agrees with other observations that the properties of the major alternative exons are similar to those of the constitutive exons (Lev-Maor *et al.*, 2007). It should be noted that the minor alternative regions seem to be the first major class of sequences, for which the positive selection could be demonstrated in the human genome.

References

Artamonova II and Gelfand MS (2004) Evolution of the exon-intron structure and alternative splicing of the MAGE-A family of cancer/testis antigens. *Journal of Molecular Evolution* **59**: 620–631.

Baek D and Green P (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proceedings of the National Academy of Sciences of the USA* **102**: 12813–12818.

Brett D, Pospisil H, Valcárcel J *et al.* (2002) Alternative splicing and genome complexity. *Nature Genetics* **30**: 29–30.

Bustamante CD, Fledel-Alon A, Williamson S *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature* **437**: 1153–1157.

Dewey CN, Rogozin IB and Koonin EV (2006) Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* **7**: 311.

Ermakova EO, Nurtdinov RN and Gelfand MS (2006) Fast rate of evolution in alternatively spliced coding regions of mammalian genes. *BMC Genomics* **7**: 84.

Fay JC, Wyckoff GJ and Wu C (2002) Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**: 1024–1026.

Gotea V and Makałowski W (2006) Do transposable elements really contribute to proteomes? *Trends in Genetics* **22**: 260–267.

Graveley BR (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics* **17**: 100–107.

Irimia M, Rukov JL, Penny D *et al.* (2008) Widespread evolutionary conservation of alternatively spliced exons in *Caenorhabditis*. *Molecular Biology and Evolution* **25**: 375–382.

Jordan IK, Wolf YI and Koonin EV (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evolutionary Biology* **4**: 22.

Kim E, Magen A and Ast G (2007) Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research* **35**: 125–131.

Kim H, Klein R, Majewski J *et al.* (2004) Estimating rates of alternative splicing in mammals and invertebrates. *Nature Genetics* **36**: 915–916.

Královicová J and Vorechovsky I (2007) Global control of aberrant splice-site activation by auxiliary splicing sequences: evidence for a gradient in exon and intron definition. *Nucleic Acids Research* **35**: 6399–6413.

Kurmangaliyev YZ and Gelfand MS (2008) Computational analysis of splicing errors and mutations in human transcripts. *BMC Genomics* **9**: 13.

Letunic I, Copley RR and Bork P (2002) Common exon duplication in animals and its role in alternative splicing. *Human Molecular Genetics* **11**: 1561–1567.

Lev-Maor G, Goren A, Sela N *et al.* (2007) The 'alternative' choice of constitutive exons throughout evolution. *PLoS Genetics* **3**: e203.

Lev-Maor G, Sorek R, Shomron N *et al.* (2003) The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science* (New York, NY) **300**: 1288–1291.

Malko D, Ermakova E and Gelfand M (2009) Evolution of structure and sequence in alternatively spliced *Drosophila* genes. Proceedings of the fourth International Moscow Conference on Computational Molecular Biology MCCMB'09, Moscow, Russia, July 20–23.

Malko DB, Makeev VJ, Mironov AA *et al.* (2006) Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes. *Genome Research* **16**: 505–509.

McDonald JH and Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**: 652–654.

McGlincy NJ and Smith CWJ (2008) Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends in Biochemical Sciences* **33**: 385–393.

Modrek B and Lee CJ (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genetics* **34**: 177–180.

Ner-Gaon H, Leviatan N, Rubin E *et al.* (2007) Comparative cross-species alternative splicing in plants. *Plant Physiology* **144**: 1632–1641.

Nurtdinov RN, Artamonova II, Mironov AA *et al.* (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Human Molecular Genetics* **12**: 1313–1320.

Nurtdinov RN, Neverov AD, Favorov AV *et al.* (2007) Conserved and species-specific alternative splicing in mammalian genomes. *BMC Evolutionary Biology* **7**: 249.

Parmley JL and Hurst LD (2007) Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Molecular Biology and Evolution* **24**: 1600–1603.

Ramensky VE, Nurtdinov RN, Neverov AD *et al.* (2008) Positive selection in alternatively spliced exons of human genes. *American Journal of Human Genetics* **83**: 94–98.

Sabeti PC, Schaffner SF, Fry B *et al.* (2006) Positive natural selection in the human lineage. *Science* (New York, NY) **312**: 1614–1620.

Sabeti PC, Varilly P, Fry B *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**: 913–918.

Severing EI, van Dijk AD, Stiekema WJ *et al.* (2009) Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. *BMC Genomics* **10**: 154.

Smith NGC and Eyre-Walker A (2002) Adaptive protein evolution in *Drosophila*. *Nature* **415**: 1022–1024.

Sorek R, Ast G and Graur D (2002) Alu-containing exons are alternatively spliced. *Genome Research* **12**: 1060–1067.

Sorek R, Shamir R and Ast G (2004) How prevalent is functional alternative splicing in the human genome? *Trends in Genetics* **20**: 68–71.

Stamm S, Ben-Ari S, Rafalska I *et al.* (2005) Function of alternative splicing. *Gene* **344**: 1–20.

Wang ET, Sandberg R, Luo S *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476.

Xing Y and Lee C (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proceedings of the National Academy of Sciences of the USA* **102**: 13526–13531.

Xing Y, Wang Q and Lee C (2006) Evolutionary divergence of exon flanks: a dissection of mutability and selection. *Genetics* **173**: 1787–1791.

Zhang XH and Chasin LA (2006) Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proceedings of the National Academy of Sciences of the USA* **103**: 13427–13432.

Zheng CL, Fu X and Gribskov M (2005) Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA* (New York, NY) **11**: 1777–1787.

Further Reading

Artamonova II and Gelfand MS (2007) Comparative genomics and evolution of alternative splicing: the pessimists' science. *Chemical Reviews* **107**: 3407–3430.

Ast G (2004) How did alternative splicing evolve? *Nature Reviews. Genetics* **5**: 773–782.

Boue S, Letunic I and Bork P (2003) Alternative splicing and evolution. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* **25**: 1031–1034.

Koonin EV (2006) The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biology Direct* **1**: 22.

Lareau LF, Green RE, Bhatnagar RS *et al.* (2004) The evolving roles of alternative splicing. *Current Opinion in Structural Biology* **14**: 273–282.

Martin W and Koonin EV (2006) Introns and the origin of nucleus-cytosol compartmentalization. *Nature* **440**: 41–45.

Roy SW and Gilbert W (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Reviews. Genetics* **7**: 211–221.

Xing Y and Lee C (2006) Alternative splicing and RNA selection pressure – evolutionary consequences for eukaryotic genomes. *Nature Reviews. Genetics* **7**: 499–509.