BIOINFORMATICS DISCOVERY NOTE Vol. 27 no. 22 2011, pages 3079–3084
doi:10.1093/bioinformatics/btr532

Genome analysis

Advance Access publication October 12, 2011

The strength of intron donor splice sites in human genes displays a bell-shaped pattern

Kai Wang[1,2,3], Rasmus Wernersson[1] and Søren Brunak[1,*]

[1]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark, [2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health and [3]Department of Radiology, Harvard Medical School, Harvard University, Boston, MA 02115, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: The gene concept has recently changed from the classical one protein notion into a much more diverse picture, where overlapping or fused transcripts, alternative transcription initiation, and genes within genes, add to the complexity generated by alternative splicing. Increased understanding of the mechanisms controlling pre-mRNA splicing is thus important for a wide range of aspects relating to gene expression.

Results: We have discovered a convex gene delineating pattern in the strength of 5′ intron splice sites. When comparing the strengths of >18 000 intron containing Human genes, we found that when analysing them separately according to the number of introns they contain, initial splice sites were always stronger on average than subsequent ones, and that a similar reversed trend exist towards the terminal gene part. The convex pattern is strongest for genes with up to 10 introns. Interestingly, when analysing the intron containing gene pool from mouse consisting of >15 000 genes, we found the convex pattern to be conserved despite >75 million years of evolutionary divergence between the two organisms. We also analysed an interesting, novel class of chimeric genes which during spliceosome assembly are fused and in tandem are transcribed and spliced into a single mature mRNA sequence. In their splice site patterns, these genes individually seem to deviate from the convex pattern, offering a possible rationale behind their fusion into a single transcript.

Contact: brunak@cbs.dtu.dk

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on November 22, 2010; revised on August 2, 2011; accepted on September 5, 2011

1 INTRODUCTION

For protein coding genes, the most significant variety spawning factor is the process responsible for intron removal (Wang *et al.*, 2008). Here the binding reaction between the highly conserved U1 snRNA 5′ terminus and the pre-mRNA intron donor splice site motif is the first and most crucial step (Freund *et al.*, 2003; Zhuang and Weiner, 1986). The consensus sequence of human donor splice sites, the sequence motif which U1 snRNA recognizes, reflects an 'average' complementarity to U1 snRNA (Lerner *et al.*, 1980; Zhuang and Weiner, 1986). The calculation of the free energy of the base pairing between the U1 snRNA 5′ terminus and donor site motif has previously been performed as an effective strategy to study binding stability and the strength of donor splice sites (Roca *et al.*, 2005; Zychlinski *et al.*, 2009).

2 METHODS

2.1 Genome data

Human genome data (Build 36) and mouse genome data (Build 36) were downloaded from the NCBI ftp site: ftp://ftp.ncbi.nlm.nih.gov/genomes/. For the extraction of both DNA sequence and annotation of intron/exon structure, we used an approach similar to FeatureExtract (Wernersson, 2005). Full-length transcripts including untranslated regions (UTRs) were extracted by paring the chromosome Contig FASTA files with the information about UTR regions and coding regions contained in the seq_genes.md file. We used the NCBI based annotation (ref) and ignored alternative annotations. Furthermore, in the few cases where the same transcript ID is present on different chromosomes (mainly some regions on the X and Y chromosomes), the first occurrence was used. Non-coding RNA genes, genes without introns and questionable/frameshift containing entries were discarded, such as genes where the sizes of the coding regions are not multiples of three, entries that begin with the prefix XM_ and XR_ which are model reference sequences produced automatically by NCBI's Genome Annotation project. In total, we extracted 18 148 reliable intron-containing Human genes and 15 318 intron-containing mouse genes all with GT–AG intron splice site pairs throughout the transcripts.

2.2 Free energy calculation

We used the MFOLD algorithm with a customized set of parameters (see below) to estimate the U1 snRNA-donor site binding energies. This approach takes into account the negative energy for stabilization of stacked base pairs and subtracts positive energy for loops, bulges and other destabilizing structures (Garland and Aalberts, 2004; Zuker, 2003). We took the conserved U1 snRNA sequence (AUACUUACCUGGC) at its 5′ terminus and joined it with the pre-mRNA section covering positions −6 to +8 relative to the donor site. This binding product, U1 snRNA:pre-mRNA, can thus be considered as a 27 bp single strand RNA fold. The last two nucleotides on U1 snRNA and the first three nucleotides on pre-mRNA were replaced by five Ns, and furthermore prevented the middle five nucleotide linkers from pairing, as well as ensured that the GT in the pre-mRNA always base pair to CA in the U1 snRNA. Then, the minimum free energy was calculated after subtracting the penalty value for the hairpin. In some cases where more than one structure could be formed based on one sequence, the optimal folding structure was selected. We tried to use different lengths of the pre-mRNA donor site motifs and U1 snRNA 5′ terminal sections as the binding products

*To whom correspondence should be addressed.

when calculating the energies (Freund *et al.*, 2005), such as the segments from position −6 to +10 and −6 to +6 on pre-mRNA, and always obtained similar results (as those displayed in Fig. 1). We concluded that the exact choice of region used is not critical to the results reported here. Here, we have chosen to use the MFOLD program; however, it should be noted that other programs use the same energy model and dynamic programming algorithm, e.g. RNAcofold from the Vienna RNA package (Hofacker, 2003; Zuker and Stiegler, 1981).

Recently, an interesting discovery of an alternative, shifted interaction between 5′ donor sites and U1 snRNAs was reported. Atypical 5′ splice sites that cannot base pair very well to U1 snRNAs actually can be recognized by U1 snRNAs efficiently by shifting one nucleotide (Roca and Krainer, 2009). However, since only 59 atypical 5′ splice sites in human were found, the new finding does not affect the genome-wide analyses reported here.

3 RESULTS

We have compared the strengths of binding associations in 18 148 intron containing Human genes and found that when analysing them separately according to the number of introns they contain, initial splice sites were always stronger on average than subsequent ones, and that a similar reversed trend exist towards the terminal gene part (Fig. 1a). In all categories representing different numbers of introns (from 1 to 18), the splice site strength display a convex, bell-shaped trend where weaker splice sites are found internally. The average strength of initial donor sites is always below −12 kcal/mol (and most often as low as −12.5 kcal/mol), while internal splice sites tend to be above.

Surprisingly, when the same analysis was repeated for 15 318 intron-containing genes in the mouse genome, the result was highly similar (Fig. 1b and Supplementary Table S1), despite the fact that most 5′ splice sites differ in sequence, while the other binding partner, the human and mouse U1 motif, is identical. Moreover, tissue-specific transcriptional regulation and splicing has diverged considerably between human and mouse. For example, 41–89% of *cis*-regulatory regions in one species were not found in the other (Odom *et al.*, 2007), and >11% of alternatively skipped exons were species-specific (Pan *et al.*, 2005)—indicating fast change over 75 million years of evolution (Kitazoe *et al.*, 2007; Waterston *et al.*, 2002).

It should be noted that while the underlying distribution in binding energies is quite broad (Fig. 2 and Supplementary Fig. S1)—the mean is estimated with high accuracy—as can be see from Table 1 illustrating this for the human data. The general trend of internal introns having weaker binding (higher free energy) holds true even when dividing the dataset into three broad categories representing initial introns, internal introns and terminal introns separately (Fig. 2). The mean (initial: −12.37 kcal/mol; internal: −11.76 kcal/mol; terminal: −12.09 kcal/mol) is significantly different between each pair of categories ($P < 2 \times 10^{-16}$ estimated by the Welch *t*-test). Finally, one may also ask whether introns of roughly the same size also differ in splice site strength, e.g. initial versus internal. Using for example a 250 bp binning, we checked that is indeed the case for all length intervals (>2500 bp in one bin). There is a considerable margin for all categories over 250 bp in intron length, while it is smaller for short introns (which are infrequent in the Human gene pool).

A subset of the initial and terminal introns is found in untranslated regions (UTRs), where there is no protein coding selection pressure on the exon part of the splice site motifs. To assess the effect of introns in UTRs, we analyzed all Human genes only containing CDS introns and found again the exact same trend (Supplementary Figs S2 and S3). Interestingly, protein unaffected splice sites in 5′ and 3′ UTRs are on average more strongly bound to U1 snRNAs, than splice sites in protein coding regions, but the bell-shaped pattern in the pre-mRNA:U1 snRNA association is the same whether genes contain UTR introns or not (3′ donor sites are only marginally stronger, see Supplementary Figs S4 and S5). This may also explain why human UTR donor sites quite successfully can be predicted from the sequence despite the fact that there is no 'reading frame/non-coding transition' delivering a pattern which can be exploited by an algorithm as is the case for coding region donor sites (Eden and Brunak, 2004).

Another explanation for the stronger initial and terminal splice sites could be that alternative splicing would be less frequent here, and that the weaker and more frequent alternative splice sites in protein coding regions would lower the average free energy of U1:pre-mRNA association in the CDS. However, several reports indicate that the opposite in fact is the case. Alternative splicing seems to be even more frequent in the 5′ UTR or non-coding transcripts (Mironov *et al.*, 1999; Sammeth *et al.*, 2008); still we find that UTR splice sites are stronger and hence most likely more confidently recognized by the splicing machinery.

It is well known that there is a tendency for long introns in Human genes to have donor sites which conform more strongly to the consensus sequence hence having stronger U1 association (Xiao *et al.*, 2007). Another well-known tendency is that the initial intron is longer than non-initial introns on average (Bradnam and Korf, 2008). If most of the initial and terminal introns were long this could at least rationalize (but not necessarily explain) the bell-shaped trend shown in Figure 1. Based on the data presented here, we therefore calculated the median length of introns along the genes using the same specific categories as in Figure 1. As shown in Supplementary Figure S6, the median intron length is indeed longer in the 5′ part of the genes, while this is not the case for terminal introns. However, when investigating the strength of initial intron donor sites as a function of intron length we found that even short initial introns tend to be very strong (Fig. 3). Indeed, the figure shows that on average initial intron splice sites are stronger for all intron length intervals.

We also investigated the relationship with the GC content which is known to be elevated in the 5′ end of Human genes (Bernardi 1991; Bernardi 2000; Furey and Haussler 2003; Pozzoli *et al.*, 2008). Supplementary Figure S7 shows that the GC content in donor splice sites (using the 11bp region used for the free energy calculations) matches the known trend; initial splice sites have slightly higher GC content, while it is stable elsewhere for all categories of genes comprising 1–9 introns. The GC content variation in the splice sites (or the genes) does thus not explain the binding energy trend as it does not display a bell-shaped trend. Also, importantly the U1 sequence has more A/T (7 bp) than G/C (4 bp) and this means that higher GC content in itself in random sequence will have a tendency to produce weaker splice sites due to less frequent Watson–Crick base pairing. We show in Figure 1 that the opposite is the case.

From the bell-shaped pattern we have discovered, splice sites near the gene edges tend to be stronger on average, and it seems to be a signature of the boundary of genes not reflecting the position-specific

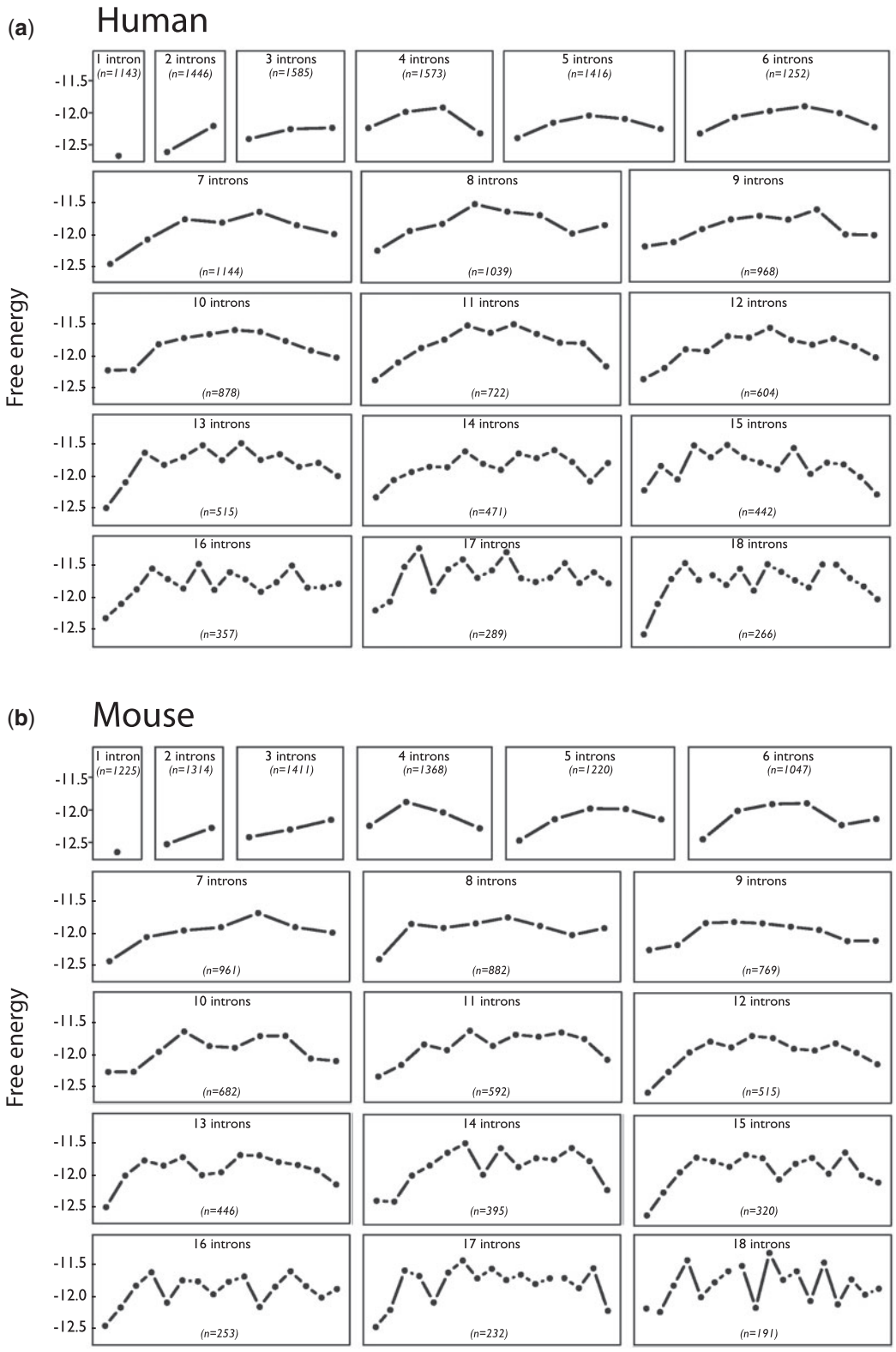

Fig. 1. Free energies of donor site pre-mRNA:U1 snRNA association in Human and mouse. (**a**) Plots showing the average energies of pre-mRNA:U1 snRNA association for introns from 16 110 Human genes with 1–18 introns. Each dot indicates the average free energy in kcal/mol (higher values means weaker splice sites). The pre-mRNA:U1 snRNA associations for initial and terminal introns have stronger binding energies, while the internal ones always have a tendency to be weak. Each panel contains information about the number of genes with that particular number of introns. Notice that as genes with an increasingly large number of introns is analysed, the sample size per panel decreases. (**b**) The same calculation for 13 854 mouse genes that contains 1–18 introns.

Table 1. Average binding energy - standard error of the mean

	Intron 1	Intron 2	Intron 3	Intron 4	Intron 5	Intron 6	Intron 7	Intron 8	Intron 9	Intron 10	Intron 11	Intron 12	Intron 13	Intron 14	Intron 15	Intron 16	Intron 17	Intron 18
1	0.078																	
2	0.070	0.070																
3	0.066	0.067	0.066															
4	0.068	0.069	0.065	0.068														
5	0.068	0.067	0.073	0.071	0.069													
6	0.076	0.074	0.076	0.080	0.076	0.078												
7	0.078	0.080	0.079	0.079	0.080	0.079	0.077											
8	0.084	0.084	0.081	0.083	0.087	0.082	0.080	0.080										
9	0.086	0.086	0.085	0.080	0.088	0.087	0.083	0.085	0.088									
10	0.088	0.091	0.088	0.092	0.094	0.094	0.094	0.089	0.090	0.091								
11	0.098	0.100	0.097	0.099	0.097	0.098	0.102	0.104	0.099	0.100	0.097							
12	0.104	0.110	0.106	0.109	0.114	0.109	0.110	0.109	0.108	0.105	0.110	0.120						
13	0.122	0.114	0.118	0.115	0.123	0.119	0.119	0.118	0.118	0.123	0.111	0.116	0.120					
14	0.125	0.111	0.123	0.122	0.126	0.122	0.112	0.124	0.115	0.114	0.121	0.118	0.119	0.125				
15	0.128	0.118	0.121	0.127	0.126	0.133	0.129	0.126	0.125	0.121	0.121	0.125	0.123	0.127	0.132			
16	0.143	0.130	0.138	0.134	0.150	0.142	0.139	0.147	0.143	0.144	0.144	0.142	0.139	0.134	0.137	0.133		
17	0.153	0.155	0.148	0.146	0.159	0.154	0.154	0.159	0.157	0.147	0.153	0.151	0.146	0.164	0.159	0.153	0.153	
18	0.160	0.173	0.164	0.166	0.163	0.161	0.163	0.170	0.170	0.171	0.145	0.162	0.168	0.172	0.170	0.162	0.163	0.159

The standard error of the mean calculated for the gene categories 1–18 (containing 1–18 introns) based on the same data as Figure 1 and Supplementary Figure S1.

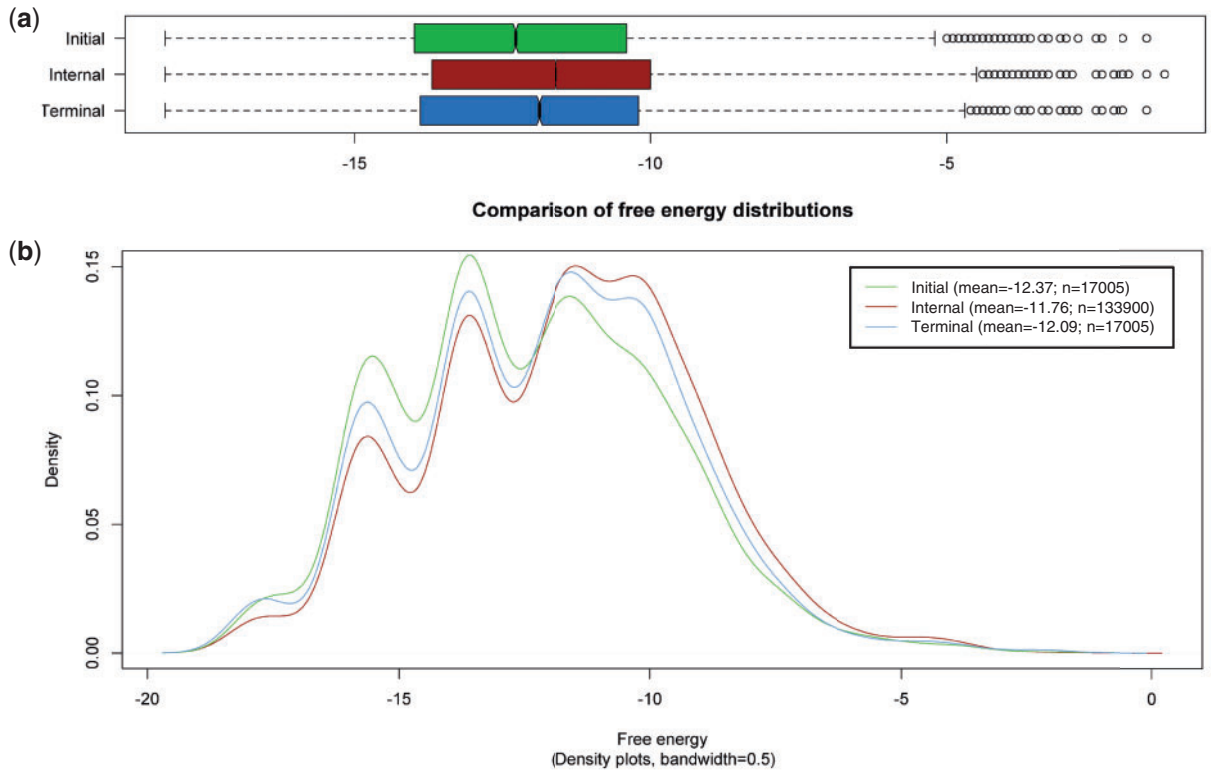

Fig. 2. Distribution of free energy in intial, internal and terminal introns. (**a**) Notched box and whiskers plots, showing the mean and distribution of the free energy for (i) all initial introns; (ii) all internal introns; and (iii) all terminal introns. Genes with a single intron was excluded from this analysis. (**b**) Density plots of the same datasets as (A).

bias in intron lengths nor GC content. At the single gene level, we also investigated how large a fraction of the individual human transcripts conformed exactly to the pattern show in Figure 1. This was done by comparing the fit of the observed free energy profile for each transcript to the background distribution expected by chance (see Supplementary Table S3 for details). We found that 7–10% of the transcripts in each category to be in an optimal fit to the profiles shown in Figure 1 at the 95% confidence level.

An interesting, novel class of genes is those which are fused and in tandem are transcribed and spliced into a single mature mRNA

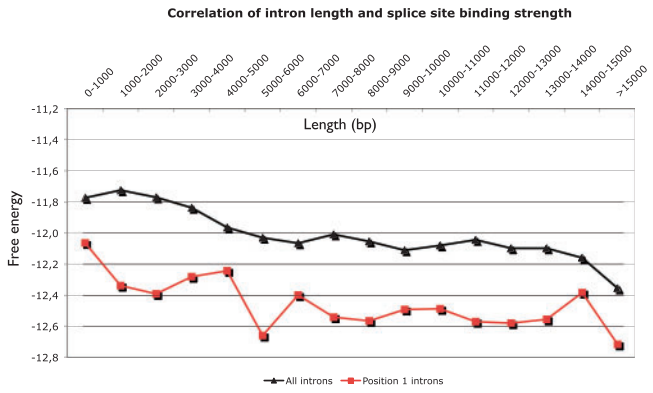

Fig. 3. Energy statistics and intron lengths. Average free energies of the pre-mRNA:U1 snRNA association compared between all introns (black curve) and initial introns (red curve).

sequence encoding a putative chimeric protein. From data generated within the ENCODE project, it has been estimated that 4–5% of the gene pairs in the Human genome can be transcribed in this chimeric manner (Akiva *et al.*, 2006; Parra *et al.*, 2006). We analysed the most abundant transcription-induced chimeras (TICs), by which two genes in tandem skip the last exon of the upstream gene and the first exon of the downstream gene. Thus, the intergenic region is removed and all other exons are joined into a single spliced mRNA (the first donor site of the downstream gene is not used). Interestingly, when analysing 46 TIC downstream genes from a recent study (Akiva *et al.*, 2006), we found that the bell-shaped pattern was indeed absent, and that the first donor sites in these genes were weaker than the general average for first donor sites (−12.45 versus −11.86 kcal/mol). Again, this was also the case when analysing the TIC genes separately according to the number of introns as in Figure 1 (except in the subset having six introns, see Supplementary Table S2). This analysis again supports the notion that the bell-shaped pattern is related to the delineation of the spliced part of the gene transcript.

4 DISCUSSION

In this analysis, we have used the U1-donor splice sites motif complementarity to assess the strength of a particular splice site. However, many other factors affect tissue-specific splicing *in vivo* including splicing enhancers, silencers, trans-acting factors, acceptor site motifs, branch point sequences, RNA secondary structure, nucleosome, chromatin and isochore features (Black, 2003; Castle *et al.*, 2008; Goren *et al.*, 2010; Luco *et al.*; Tilgner *et al.*, 2009). Given this additional complexity, it is not surprising that we could not find the bell-shaped pattern present in every single Human and mouse gene just by investigating one of the contributing factors, the U1-donor site association. Different splicing regulatory elements compensate for weaker splice sites in a coevolutionary network (Xiao *et al.*, 2007). Additionally, as already mentioned above the size of exons and introns correlates with the strength of splice sites and there is no obvious way to quantitatively assess or make compensation for all these factors in the analysis presented here (see Supplementary Fig. S8 for statistics involving splicing enhancers).

The conclusion from this analysis is that spliced Human genes seem to contain on average a 'delineation trend' in the donor splice sites, which may play a key role in spliceosome recognition. This role may be most significant in the set of genes where the trend is present at the level of single genes (∼10% of the gene pool), or it may go far beyond that depending on the contributions of the many other splicing site strength influencing factors mentioned above. Despite >75 millions of years of divergence, mouse genes display the exact same trend. The discovery cannot answer the puzzle of spliceosomal intron evolution or why genes are intervened by introns, but it highlights an interesting pattern which in very different types of introns (both within coding regions and UTRs) seem to be subject to similar evolutionary pressures conserving the trend in strength. It also suggests a potential future direction for comparative analysis of spliced genes and gene structure prediction (including splice site prediction) by defining gene categories according to the number of introns they contain.

ACKNOWLEDGEMENTS

We thank Henrik Bjørn Nielsen and Agnieszka Sierakowska Juncker for useful comments.

Funding: Danish National Research Foundation; Natural Science Research Council of Denmark.

Conflict of Interest: none declared.

REFERENCES

Akiva,P. *et al.* (2006) Transcription-mediated gene fusion in the human genome. *Genome Res.*, **16**, 30–36.
Bernardi,G. (1991) CpG islands, genes and isochores in the genomes of vertebrates. *Gene*, **106,** 185–95.
Bernardi,G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene*, **241,** 3–17.
Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.
Bradnam,K.R. and Korf,I. (2008) Longer first introns are a general property of eukaryotic gene structure. *PLoS One*, **3**, e3093.
Castle,J.C. *et al.* (2008) Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat. Genet.*, **40**, 1416–1425.
Eden,E. and Brunak,S. (2004) Analysis and recognition of 5′ UTR intron splice sites in human pre-mRNA. *Nucleic Acids Res.*, **32**, 1131–1142.
Freund,M. *et al.* (2003) A novel approach to describe a U1 snRNA binding site. *Nucleic Acids Res.*, **31**, 6963–6975.
Freund,M. *et al.* (2005) Extended base pair complementarity between U1 snRNA and the 5′ splice site does not inhibit splicing in higher eukaryotes, but rather increases 5′ splice site recognition. *Nucleic Acids Res.*, **33**, 5112–5119.
Furey,T.S. and Haussler,D. (2003) Integration of the cytogenetic map with the draft human genome sequence. *Hum. Mol. Genet.*, **12,** 1037–1044.
Garland,J.A. and Aalberts,D.P. (2004) Thermodynamic modeling of donor splice site recognition in pre-mRNA. *Phys. Rev. E*, **69**, 041903–041907.
Goren,A. *et al.* (2010) Overlapping splicing regulatory motifs–combinatorial effects on splicing. *Nucleic Acids Res.*, **38,** 3318–3327.
Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
Kitazoe,Y. *et al.* (2007) Robust time estimation reconciles views of the antiquity of placental mammals. *PLoS One*, **2**, e384.
Lerner,M.R. *et al.* (1980) Are snRNPs involved in splicing. *Nature*, **283**, 220–224.
Luco,R.F. *et al.* (2010) Regulation of alternative splicing by histone modifications. *Science*, **327**, 996–1000.
Mironov,A.A. *et al.* (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.
Odom,D.T. *et al.* (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.*, **39**, 730–732.

Pan,Q. *et al.* (2005) Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet.*, **21**, 73–77.

Parra,G. *et al.* (2006) Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res.*, **16**, 37–44.

Pozzoli,U. *et al.* (2008) Both selective and neutral processes drive GC content evolution in the human genome. *BMC Evol. Biol.*, **8,** 99.

Roca,X. and Krainer,A.R. (2009) Recognition of atypical 5′ splice sites by shifted base-pairing to U1 snRNA. *Nat. Struct. Mol. Biol.*, **16**, 176–182.

Roca,X. *et al.* (2005) Determinants of the inherent strength of human 5′ splice sites. *RNA*, **11**, 683–698.

Sammeth,M. *et al.* (2008) A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.*, **4**, 14.

Tilgner,H. *et al.* (2009) Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.*, **16**, 996-1001.

Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Waterston,R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Wernersson,R. (2005) FeatureExtract - extraction of sequence annotation made easy. *Nucleic Acids Res.*, **33**, W567–W569.

Xiao,X.S. *et al.* (2007) Coevolutionary networks of splicing cis-regulatory elements. *Proc. Natl Acad. Sci. USA*, **104**, 18583–18588.

Zhuang,Y. and Weiner,A.M. (1986) A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. *Cell*, **46**, 827–835.

Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.

Zychlinski,D. *et al.* (2009) Limited complementarity between U1 snRNA and a retroviral 5′ splice site permits its attenuation via RNA secondary structure. *Nucleic Acids Res.*, **37**, 7429–7440.