INTRON-EXON PATTERNS AS A POTENTIAL TOOL IN STUDYING GENE EVOLUTION

A. Ruvinsky

University of New England, Armidale NSW 2350, Australia, e-mail: aruvinsk@une.edu.au

The majority of introns are ancient elements and their phases and positions in genes were preserved for a long time. A string of intron phases represents a structure which carries essential information about organization and evolution of genes, which is usually ignored. Numerous observed strings have non-random intron phase patterns caused by intragenic repeats. Correlation between the lengths of CDS and the number of introns per human gene is high. Lengths of exons often remain constant in homologous and even paralogous genes belonging to distant species. Alignment of exon-intron strings provides useful visualization and generates new knowledge about evolution of gene families. It unravels intragenic duplications, intron gains and losses as well as extensions and contractions of exons. This additional information seems to be useful for studying gene evolution.

Key words: intron, exon, alignment, intragenic duplications, gene families

Introduction

Positions and phases of the majority of introns show a great deal of conservation (Rogozin *et al*., 2003; Roy, Gilbert, 2005). There are 3 phases, in which introns can be inserted: between codons (phase 0) and after the first or second nucleotides of a codon (phases 1 & 2). Shifts of intron–exon boundaries changing intron phases are rare events and have limited effect on the overall picture (Rogozin *et al.*, 2000). Intron gains and losses are more frequent and they certainly affect exon-intron structures of genes but do not necessarily influence corresponding proteins. Intragenic duplications likely played an important role in evolution of some genes (Jacob, 1983; Li, 1983; Patthy, 1987). According to available estimates the proportion of duplicated exons in long human genes is at least 6 % (Fedorov *et al*., 1998) and duplicated sequences occur in about 14 % of all proteins (Marcotte *et al*., 1999). There are hundreds of highly redundant genes in the human genome (Ruvinsky, Watson, 2007) and frequency of internal duplications has been increasing during metazoan evolution (Chen *et al.*, 2007). Intron-exon patterns allow tracing past events and could be helpful in evolutionary reconstructions. For example, a string of intron phases, like 0112121111111211211211121112111, representing a structure of human *GTF2I* gene, coding for general transcription factor 2I, contains valuable data. Three genes were identified in this family (Makeyev *et al*., 2004). Lengths of exons which in some cases remains stable for lengthy evolutionary periods is another useful source of information. More detailed analysis of *GTF2I* gene confirms presence of several intragenic duplications and sheds light on the evolution of the gene. Those genes, which are prone to internal duplications, eventually became lengthy and their evolutionary pathways could be affected. Duplications involving an exon and sections of surrounding introns or several exon-intron pairs, if they framed by introns in the same phase, do not affect reading frame as well as exon lengths. Alignments of exon-intron structures of several genes from different species belonging to the same gene family could provide valuable information. This approach may help discriminate orthologs and paralogs and show the differences in evolutionary pathways of genes, including losses and gains of exons and introns and other intragenic rearrangements. The challenge is to understand the reasons behind these changes.

Materials and Methods

The data was extracted from the exon-intron database (Saxonov *et al*., 2000), which was extensively purged. The longest of the duplicate genes were left in the database and considered the constitutive form. The total numbers of studied genes were: *Hs*-11,315, *Dm*-8,497, *Ce*-10,312 and *At*-9,914. Some information was also obtained from genome browser Ensembl (http://www.ensembl.org/index.html) Statistical analysis was performed using methods described in our recent publication (Ruvinsky, Watson, 2007).

Results

Comparisons of entropy values between observed intron strings and randomly simulated in Bernoulli schemes revealed that numerous observed strings have non-random intron phase patterns. The frequency of outliers among human genes which are beyond $Z_{2.58}$ threshold (0,01 of the normal distribution) is 3,2 times higher than expected and is getting much higher for stricter Z thresholds (Ruvinsky, Watson, 2007). Many of such outliers have intragenic repeats. Correlation between the lengths of CDS and the number of introns per human gene is high ($r = 0,83$) and getting stronger as number of introns increases. A possible interpretation of this fact is that intragenic duplications are more frequent in the genes with numerous introns and, because exons are also parts of the duplications, the length of coding sequence stronger correlates with introns number. Recently Chen *et al*. (2007) came to a comparable conclusion studying repeats in proteins. *GTF2I* is an example of a human gene with several intragenic repeats (Fig. 1).

Highly conservative exons located in the middle of these 6 repeats show significant DNA sequence similarity and hence the origin from a common ancestral sequence. All these 6 repeated exons have exactly the same length, there is no sequence gap in any of them and there are many conservative positions. The level of sequence identity varies from 66 % to ~ 40 % in 184 nucleotides. The total number of duplication events is likely to be 5. Identity of amino acid sequences coded by the conservative exons varies from 66.7 to 38,3 % and they belong to a highly conserved domain (pfam02946.12.) with DNA binding function (Vullhorst, Buonanno, 2005). Alignment of exon-intron structures of genes from *GTF2I* family from several vertebrate species (Table 1) shows a great deal of conservation particularly between *GTF2I* orthologs from *Homo sapiens*, *Gallus gallus* and *Xenopus tropicalis*. Three other orthologs (*GTF2IRD1*) from fish species *Danio rerio* and *Oryzias latipes* and *Takifugu rubripes*, being paralogs to the tetrapod genes, show both similarities and differences in exon-intron structure. Intron insertions are likely the cause of the steadily increasing number of exons between the first and the second repeats. The fish species have only one lengthy exon following the first GTF2I repeat, while in frogs there are 5 exons, in birds 6 and in mammals 7, all of which are rather short. Intron loss, on the contrary, is a plausible explanation for the existence 268 nucleotides exons in fish species. The corresponding position of the gene in other compared vertebrate species contain two exons of 68 and 184 nucleotides, total of which is equal to 268. Taking into consideration that the 184-nucleotide exon is an ancient element in this gene family and surrounding introns are in the same phases, more parsimonious assumption is loss of the intron in the common ancestor of fish species. An alternative explanation based on insertion of phase 1 intron in higher vertebrates seems unlikely. Comparisons of exon-intron structures also show shifts of reading frames. For instance, shifting exon-intron boundary can be observed in *Xenopus tropicalis* 33 nucleotides exon (Table 1, underlined exon). It differs from the corresponding exons in other species by 4 extra nucleotides, such addition must change phase of the following intron from 1 to 2. This expectation is matched by the observation. The *GTF2IRD1* genes from fish species also contain modified repeat at the 3' end, which has length of 193 nucleotides (184 + 9) and thus has 3 extra codons. This is another example of exon expansion. The tetrapod species also have *GTF2IRD1* genes (not shown at Table 1), which are very similar to the fish species. However, *GTF2I* orthologs are not known for the fish species.

Discussion

Intragenic duplications can, at least in some degree, explain creation of introns and exons. Studies of protein families revealed distinct duplication patterns and improved current understanding of

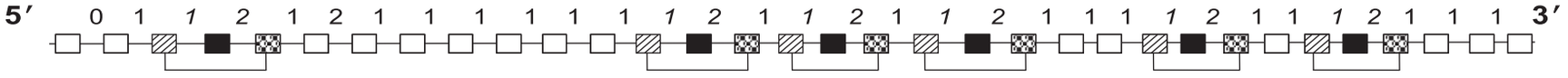

Fig. 1. Schematic presentation of exon-intron structure of human gene coding for general transcription factor 2I, *GTF2I*.

Numbers indicate intron phases. Rectangulars represent exons and lines introns. The brackets below the figure show the areas of repeats, which include highly conservative exons (black rectangulars), as well preceding (striped rectangulars) and the following exons (rectangulars with balls) and introns (clear) between them. Conservative exons within the repeats are framed by introns in phase 1 and 2 (italic). Numeration of exons (not shown) starts from the first protein coding exon at the 5' end and finishes with the last protein coding exon at the 3' end.

Table 1

Alignment of exon lengths and intron phases for genes from GTF2I family in several vertebrate species

Species*	Length in aa	Number of exons	Number of GTF2 repeats	Exon lengths (above) and intron phases (below)
Hs	998	33	6	104,139,135, 184, 29,55,44,78,60,57,63,111,66,184,59,72,184,59,72,184,59,75,102,66,184,56, 81,84,184,29,42,42,76 0 1 1 2 1 2 1 1 1 1 1 1 1 2 1 1 2 1 1 2 1 1 1 1 2 1 1 1 2 1 1 1
Gg	984	32	6	106,139,135, 184, 29,49,44,78,60, ,63,102,66,184,59,72,184,59,72,184,59,84, 96,66,184,53,111,84,184,29,69,15,76 0 1 1 2 1 2 1 1 1 1 1 1 2 1 1 2 1 1 2 1 1 1 1 1 2 1 1 1 2 1 1 1
Xt	930	31	6	96,139,123, 184, 29,43,44,78,54, 96,66,184,56,72,184,59,72,184,59,63, 93,66,184,53, 78,84,184,33, 5,48,80 0 1 1 2 1 2 1 1 1 1 1 2 1 1 2 1 1 2 1 1 1 1 1 2 1 1 1 2 2 1 1
Dr	919	26	5	126,146,140, 184,225, 96,84,184,26,130,23,81,37,23,51, 66,184,62, 39, *268*,29, 78,193,38,116,66 0 2 1 2 1 1 1 2 1 2 1 1 2 1 1 1 2 1 1 2 1 1 2 1 0
Ol	958	27	5	126,142, 63,174, 184,269, 90,84,184,26,124,98,30,36,55,43, 40,184,62, 78, *268*,29, 78,193,38,110,69 0 1 1 1 2 1 1 1 2 1 2 1 1 1 2 0 1 2 1 1 2 1 1 2 1 0
Tr	930		5	45 18 45, 186,65,231,317, 87,84,184,26,130,116,126, 90,184,62, 72, *268*,29, 78,193,38,110,69 0 0 0 0 0 2 2 1 1 1 2 1 2 1 1 1 2 1 1 2 1

3 *GTF2I* orthologs from the tetrapods and 3 *GTF2IRD1* orthologs from fish species (tinted) represent two paralogous groups of genes from the same family. Exon lengths are shown above and intron phases below. The alignment of exon-intron strings was confirmed by DNA sequence similarity. Exons with length 268 are italicized and exons with length of 33 and 193 nucleotides are underlined. Explanations are in the text.

* *Homo sapiens – human*; *Gallus gallus* – chicken; *Xenopus tropicalis* – frog; *Danio rerio* – zebrafish; *Oryzias latipes* – medaka; *Takifugu rubripes* – pufferfish.

the process. Tandem repeats of certain domains can be observed in many proteins (Björklund *et al*., 2006). A model of gene formation based on essential role of introns in the duplication process was recently suggested (Street *et al*., 2006). Similar observation relevant to MHC-linked *tenascin-X* gene has been earlier made by Hughes (1999). Our data support the view that intragenic duplications were used extensively during evolution of lengthy genes. Symmetric exons or clusters of neighbouring exons framed by introns in the same phase are preferable for duplication process (Long *et al*., 1998). If the breaks occur in the surrounding introns, which are inserted in the same phase, this does not shift the reading frame and might not cause negative consequences. As we observed, several consecutive duplications create highly repetitive intron strings detectable by measuring their entropy.

A combined search for exons of the equal length framed by introns in the same phase suggested here is the efficient approach for finding intragenic duplications. Finally such intragenic duplications, involving a single exon-intron pair or more complex grouping, can be confirmed by the alignments of DNA and protein sequences. Long genes resulted from numerous internal duplications are not very common, but could become important if their proteins became «hubs» of proteome interactions (Dosztányi *et al*., 2006). In some cases considered in this paper, intragenic repeats have a tandem structure, which might be a product of unequal recombination. In other situations intragenic repeats are dispersed. The basic point, however, remains unchanged, intragenic repeats regardless of their lengths or positions have to be framed by introns in the same phase. This is an essential condition for successful unequal recombination; otherwise shift of reading frame is inevitable.

Alignments of exon-intron structures from the same gene family may provide useful information, which can add to classical methods of DNA and protein sequences comparisons. Easy visualization of very lengthy alignments is the obvious advantage. It also can be helpful in distinction between orthologous and paralogous genes from the same family, because it utilises information about intron phase distribution and exon length never used by the standard methodology. Lastly, the alignments of exon-intron structures provide a wealth of new knowledge about all kinds of intragenic rearrangements, including intron gains and losses, exon expansions and contractions as well as other changes, which should bring additional opportunities for reconstruction of gene evolution.

Acknowledgement

The author is grateful to Dr. C. Watson for his contribution to the original paper «Intron phase patterns in genes: preservation and evolutionary changes» published in The Open Evolution Journal 2007, 1, 1–14.

References

Björklund Å.K., Ekman D., Elofsson A. Expansion of protein domain repeats // PLoS Computational Biol. 2006. 2. P. 0959–0970.

Chen C.-C., Li W.-H., Sung H.-M. Patterns of internal gene duplication in the course of metazoan evolution // Gene. 2007. 396. P. 59–65.

Dosztányi Z., Chen J., Dunker A.K. *et al*. Disorder and sequence repeats in hub proteins and their implications for network evolution // J. Proteome Res. 2006. 5. P. 2985–2995.

Fedorov A., Fedorova L., Starchenko V. et al. Influence of exon duplication on intron and exon phase distribution // J. Mol. Evol. 1998. 46. P. 263–271.

Hughes A.L. Concerted evolution of exons and introns in the MHC-linked tenascin-X gene in mammals // Mol. Biol. Evol. 1999. 16. P. 1558–1567.

Jacob F. Molecular tinkering in evolution. Cambridge: Cambridge Univ. Press, 1983.

Li W.-H. Evolution of duplicate genes and pseudogenes // Evolution of Genes and Proteins / Eds M. Nei, R.K. Koehn. Sunderland, MA.: Sinauer Associates Inc, 1983. P. 14–37.

Long M., de Souza S., Rosenberg C., Gilbert W. Relationship between “proto-splice sites” and intron phases: evidence from dicodon analysis // Proc. Natl Acad. Sci. USA. 1998. 95. P. 219–223.

Makeyev A.V., Erdenechimeg L., Mungunsukh O. *et al*. GTF2IRD2 is located in the Williams-Beuren syndrome critical region 7q11.23 and encodes a protein with two TFII-I-like helix-loop-helix repeats // Proc. Natl Acad. Sci. USA. 2004. 101. P. 11052–11057.

Marcotte E.M., Pellegrini M., Yeates T.O., Eisenberg D. A census of protein repeats // J. Mol. Biol. 1999. 293. P. 151–160.

Patthy L. Intron-dependent evolution: preferred types of exons and introns // FEBS Lett. 1987. 214. P. 1–7.

Rogozin I.B., Lyons-Weiler J., Koonin E.V. Intron sliding in conserved gene families // Trends in Genet.

2000. 16. P. 430–432.

Rogozin I.B., Wolf Y.I., Sorokin A.V. *et al*. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution // Curr. Biol. 2003. V. 13. № 17. P. 1512–1517.

Roy S.W., Gilbert W. Rates of intron loss and gain: implications for early eukaryotic evolution // Proc. Natl Acad. Sci. USA. 2005. 102. P. 5773–5778.

Ruvinsky A., Watson C. Intron phase patterns in genes: preservation and evolutionary changes // The Open Evol. J. 2007. 1. P. 1–14.

Saxonov S., Daizadeh I, Federov A., Gilbert W. EID: the Exon-Intron Database – an exhaustive database of protein-coding intron-containing genes // Nucl. Acids Res. 2000. 28. P. 185–190.

Street T.O., Rose G.D., Barrick D. The role of introns in repeat protein gene formation // J. Mol. Biol. 2006. 360. P. 258–266.

Vullhorst D., Buonanno A. Multiple GTF2I-like repeats of general transcription factor 3 exhibit DNA binding properties: evidence for a common origin as a sequence-specific DNA interaction module // J. Biol. Chem. 2005. 280. P. 31722–31731.