Investigation of exon-intron structure multiple alignments.

I.V. Poverennaia,

Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia, ipoverennaya@gmail.com

T.V. Astakhova, M.A. Roytberg,

Institute of Mathematical Problems in Biology RAS, Pushchino, Russia, Higher School of Economics, Moscow, Russia, astakhova@lpm.org.ru, mroytberg@lpm.org.ru

Intron-rich organisms (especially vertebrates) tends to have multiple long introns. For instance, approximately 90% and 40% of genes in primate genomes contain at least one intron with length more than 1 kbp and 10 kbp, respectively. The existence of extremely long introns have been of interest for a long time and different studies about intron length have been done. Long introns seem to evolve faster than short ones; some of them might contain various regulatory elements; a positive correlation between intron length and gene expression have been shown for genes expressed at low to medium levels (highly expressed genes usually have shorter introns) [1,2].

Introns can be divided into three phases (0, 1, and 2) depending on their position relatively to the reading frame. Phase 0, 1, and 2 introns are located, respectively, before the first, after the first, and after the second nucleotide of a codon. The ratio of the three phases used to be called "golden' due to its observed constancy at 5:3:2 in almost all analyzed genomes (in some articles ratio is 2:1:1, the key point is over-representation of phase 0) [3]. Previously we have shown that the "golden ratio" does not hold if only long introns are considered. Here we present the further results. To provide easy access to analyzed data we have constructed a database containing information about intron-exon structure of 24 eukaryotic genomes (including mammals, fishes, plants, etc).

For each of five genomes (*Homo sapiens*, *Macaca mulatta*, *Mus musculu*, *Anolis carolinensis* and *Danio rerio*) we have made a set consisting of 100 genes that have a significant number of long introns and phase 1 introns. Then we have clustered the selected genes into the functional groups and selected for the primary analysis two genes *ptprd* (protein tyrosine phosphatase, receptor type, D) and *bcas3* (breast carcinoma amplified sequence 3) that are presented in all five sets. Interestingly, both genes seem to be associated

with cancer. To expand our datasets we have selected 90 orthologues for each gene from NCBI Annotation Pipeline and then reconstructed a multiple alignment of exon-intron structure for each gene set; the obtained alignments are based on comparison of exon lengths and then checked by sequence alignment.

E.g. for *ptprd* family the alignment has revealed 35 aligned "exon groups"; in a particular genome an exon group mainly consists of one exon, but may contain more than one exon or be empty at all. Correspondingly, we have obtained 34 groups of orthologous introns; each group contains from 57 to 87 introns (78 introns in average). There are different ranges of average intron lengths in different groups, see Table 1.

Table 1. Histogram of average intron lengths in different intron groups of *ptprd* genes.

Average intron length, bp	400-1000	1000-2000	2000-4000	4000-8000	8000 - 16000	over 50000
Number of intron groups	7	7	7	6	5	2

We have compared the phases and lengths of the aligned introns. The phases seem to be highly conserved; only a few changes have been found. It would be interesting to obtain more data and to analyze frequencies of the different mutations leading to the phase change. Naïve hypothesis is that the changes mostly lead to appearance of phase 0 introns but it has to be checked.

Surprisingly, the intron lengths are also conserved but in a special way. The intron lengths within a group might vary significantly, especially for long introns. However, it is changed if instead of intron length L we will consider a normalized length $N = (L-A)/A$, where A is an average length within a group of orthologous introns. E.g. in case of birds (28 species have been considered) the normalized length is beyond the interval (-0.15, 0.15) only for 14.8% of *ptprd* introns. And only for 3.2% of introns it is beyond the interval (-0.5, 0.5).

1) Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. *Bergman CM, Kreitman M. Genome Res. 2001 Aug; 11(8):1335-45.*

2) A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes. *Carmel L, Koonin EV Genome Biol Evol. 2009 Sep 22; 1():382-90.*

3) Phase distribution of spliceosomal introns: implications for intron origin. *Nguyen HD, Yoshihama M, Kenmochi N. BMC Evol Biol. 2006;6:69. doi: 10.1186/1471-2148-6-69*