# Dependence between lengths and phases of introns

T.V. Astakhova,

*Institute of Mathematical Problems in Biology, Pushchino, Russia,* astakhova@lpm.org.ru

I.I. Tsitovich,

*Institute of Information Transmission Problem, Moscow, Russia, Higher School of Economics, Moscow, Russia,* cito@iitp.ru

V.V. Yacovlev, M.A. Roytberg,

*Institute of Mathematical Problems in Biology, Pushchino, Russia, Higher School of Economics, Moscow, Russia,* mroytberg@lpm.org.ru, v.yacovlev@gmail.com

We have systematically studied dependencies between lengths, phases and positions of introns in 17 complete genomes of various taxa. Relation between intron lengths and various factors were already studied in a number of papers, see e.g. [1-3], and review [4]. However the dependencies between lengths and phases were not studied yet. Our investigation allowed both to reveal new significant regularities and to give new interpretation for some known ones. The following genomes were analyzed: *Apis mellifera, Drosophila melanogaster, Nasonia vitripennis, Tribolium castaneum, Danio rerio, Xenopus tropicalis, Anolis carolinensis, Gallus gallus, Meleagris gallopavo, Taeniopygia guttata, Mus musculus, Canis lupus familiaris, Sus scrofa, Callithrix jacchus, Macaca mulatta, Pan troglodytes. Homo sapiens.* We have studied separately 1$^{st}$ introns, internal introns and last introns. The results presented below are related to internal introns. The results are basically same for other classes, although first introns tend to be longer than other introns. We have studied histograms of intron lengths both for all introns of a species or a taxon, and depending on phases of intron itself, and of its previous and next introns.

Firstly, we have shown that histograms of intron length for *D. rerio, X. tropicalis, A. carolinensis, G. gallus, M. gallopavo, T. guttata* have a “tableland”, where frequencies of introns of given length are almost constant, see Table 1. The effect for *D. rerio* was reported in [3]. The histograms of insects and mammals do not contain such a tableland.

Secondly, we have revealed the growth of the part of introns of phase 1 and decrease of part

| №№ | SPECIES | Range of tableland | Min value of the histogram | Max value of the histogram | All inrons | Table region introns | %%Table region introns |
|---|---|---|---|---|---|---|---|
| 1 | D. rerio | 940 - 2030 | 0.15% | 0.18% | 114893 | 21207 | 18.46% |
| 2 | X. tropicalis | 300 - 750 | 0.05% | 0.05% | 80533 | 17863 | 22.18% |
| 3 | A.carolinensis | 160 - 910 | 0.03% | 0.04% | 73632 | 19895 | 27.02% |
| 4 | G.gallus | 170 - 610 | 0.60% | 0.66% | 81916 | 23105 | 28.21% |

Table 1. Table regions within histograms of intron lengths. The data for *M. gallopavo* and *T.guttata* are close to those for *G. gallus.*

of introns of phase 0 with intron length. The effect was shown for all considered species. E.g. introns of phase 1 comprise 31% of all mammalian introns, 33% within mammalian introns of length > 5000 (Z-score 12.1) and 36% within mammalian introns of length > 20000 (Z-score 14.5). Another found effect is that a neighbor of a long (short) intron tends also to be tends to be also long (short). E.g. let long introns be introns of length more than 1500 bp. Then human genome contains 51% of short introns, the empirical probability to find short intron after another short intron is 65% and to find short intron after two short introns is 75%. For long introns the corresponding values are 49%, 62.5% and 68%. Similar effects can be demonstrated for other species and cutoffs. For example, for *D.melanogaster* and intron length cut-off 150 bp the empirical probabilities for short introns are 71.5% (part of short introns), 80% (probability to find short intron after another short intron) and 85% (probability to find short intron after two short introns) and correspondingly 28.5%, 45% and 54% for long introns. Along with the above observation that long introns tend to have phase 1, this is in accordance with the effect of chains of symmetric exons of phase 1 presented in [1].

Thus we have revealed three new effects related to intron lengths and phases. The next step is explanation of the effects based on theory of intron evolution and mechanisms of splicing.


1. M.Long et al (1995) *PNAS*,**92**:12495-9.
2. I.L.Shchepetkova, M.S.Gelfand (1997) *Biofizika (rus)*, **42**:82-89.
3. S. P. Moss et al. (2011) *Genome Biol. Evol.*, **3**:1187–1196
4. I.B Rogozin, et al. (2012) *Biology Direct*, **7:**11