Dependence between lengths and phases of introns

T. V. Astakhova [1], I. I. Tsitovich [2, 3], V. V. Yacovlev [1, 3], M. A. Roytberg [1, 3]

1) Institute of Mathematical Problems in Biology, Pushcino, Russia; 2) Institute of Information Transmission Problem, Moscow, Russia; 3) Higher Scool of Economics, Moscow, Russia mroytberg@lpm.org.ru

SUMMARY

We have studied regularities in exon-intron structure in 17 complete genomes of various taxa related to intron lengths. Unlike to regularities related to intron phases, these regularities have not been carefully studied yet. The main results are as follows.

1. Part of introns of phase 1 increases and part of introns of phase 0 decreases with intron length. The effect was shown for all considered species. E. g. introns of phase 1 comprise 31% of all introns of H.sapience, 33% within introns of length > 5000 (Z-score 12.1) and 36% within introns of length > 20000 (Z-score 14.5).

2. A neighbor of a long (short) intron tends also to be long (short) intron. The effect was demonstrated for various species and cutoffs. E.g. for human genome and cutoff 1500 bp, the 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%. 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 presented of phase 1 presented in [Long M et al (1995) PNAS,92:12495-9.].

3. Histograms of intron length do not decrease monotonically but contain "tableland regions" of two types: "close" region is situated close to the peak of the histogram and is relatively short (up to 100 bp), "far" region occupies intron lengths from several hundreds bp and is much longer. Different genomes may contain one of tableland regions or both.

THE GENOMES

The table shows a list of analyzed genomes and characteristics of their introns

№	Species	Class	Number of introns	Number of genes	Average number of introns per gene
1	Apis_mellifera	Insecta	45062	7231	6,23
2	Drosophila_melanogaster	Insecta	37817	7612	4,97
3	Nasonia_vitripennis	Insecta	49014	8171	6,00
4	Tribolium_castaneum	Insecta	37843	6982	5,42
5	Danio_rerio	Osteichthyes	152162	17539	8,68
6	Xenopus tropicalis	Amphibia	106488	12170	8,75
7	Anolis_carolinensis	Reptilia	94900	10202	9,30
8	Gallus_gallus	Aves	104878	10915	9,61
9	Meleagris_gallopavo	Aves	65897	6849	9,62
10	Taeniopygia_guttata	Aves	79129	8137	9,72
11	Mus_musculus	Mammalia	142412	15881	8,97
12	Canis_lupus_familiaris	Mammalia	119018	12158	9,79
13	Sus_scrofa	Mammalia	100849	12126	8,32
14	Callithrix_jacchus	Mammalia	101376	10998	9,22
15	Macaca_mulatta	Mammalia	104132	11682	8,91
16	Pan_troglodytes	Mammalia	117872	13072	9,02
17	Homo_sapiens	Mammalia	120750	12796	9,44

LENGTHS OF ADJACENT INTRONS ARE DEPENDENT

The tables below show that an intron adjacent to a long intron tends to be long and an intron adjacent to a short intron tends to be short. The regularity take place for all thresholds and species.

Moreover, the "probability" that an intron adjacent to two long (short) introns is also long (short) is greater than the analogous probability for neighbors of one long (short) intron.

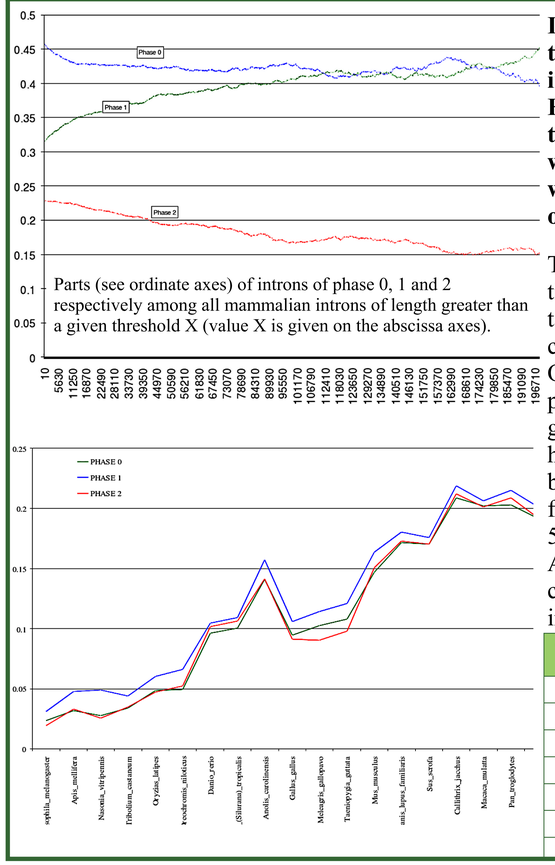
Homo Sapiens

Threshold	%Short	S->S	SS->S	S<-S	S<=SS	%Long	L->L	LL->L	L<-L	L<-LL
150	10,1%	26,2%	41,5%	25,0%	38,4%	89,9%	91,6%	92,7%	92,1%	93,5%
1000	40,2%	58,2%	71,5%	55,8%	67,5%	59,8%	71,1%	75,5%	73,1%	78,6%
1500	**51,0%**	**65,1%**	**75,0%**	**62,4%**	**70,8%**	**49,0%**	**62,5%**	**67,8%**	**65,2%**	**72,0%**
3000	70,2%	78,1%	83,2%	75,1%	79,2%	29,8%	46,4%	54,6%	50,6%	61,5%
5000	81,5%	86,7%	89,8%	83,8%	86,2%	18,5%	37,8%	48,0%	43,3%	57,7%
10000	91,0%	93,9%	95,5%	91,6%	92,6%	9,0%	31,8%	44,2%	39,6%	57,4%
20000	96,0%	97,4%	98,2%	95,9%	96,3%	4,0%	28,3%	41,1%	38,8%	54,4%
100000	99,6%	99,7%	99,8%	99,4%	99,4%	0,4%	14,9%	21,4%	25,4%	40,4%

Drosophila Melanogaster

Threshold	%Short	S->S	SS->S	S<-S	S<=SS	%Long	L->L	LL->L	L<-L	L<-LL
150	**71,52%**	**80,17%**	**84,59%**	**73,36%**	**75,08%**	**28,48%**	**44,86%**	**53,71%**	**54,44%**	**66,30%**
1000	89,53%	92,43%	93,64%	87,73%	87,24%	10,47%	26,88%	36,84%	38,55%	54,26%
1500	92,12%	94,19%	95,15%	90,03%	89,46%	7,88%	23,17%	29,77%	35,12%	48,29%
3000	95,54%	96,70%	97,31%	93,67%	93,13%	4,46%	18,82%	25,66%	31,41%	48,09%
5000	97,39%	98,01%	98,32%	95,67%	95,04%	2,61%	14,55%	19,55%	27,57%	43,21%
10000	98,90%	99,10%	99,26%	97,68%	97,19%	1,10%	9,16%	16,39%	20,93%	38,46%

LONG INTRONS and PHASE 1



Parts (see ordinate axes) of introns of phase 0, 1 and 2 respectively among all mammalian introns of length greater than a given threshold X (value X is given on the abscissa axes).

It is well known that the majority of introns have phase 0. However, we show that it is not true if we restrict ourselves with long introns only.

The effect is statistically significant and takes place for all considered taxons. One can also see that part of long introns grows from ants to humans (see Figure below, the threshold for long introns here is 5000). The excess of A. carolinensis may be caused by duplications in the genome.

Cut-off (L)	%% Phase 1	Z-score
0	31,4%	
1000	31,8%	5,9
1500	32,1%	8,7
3000	32,5%	10,6
5000	33,1%	12,1
10000	34,2%	14,1
20000	35,7%	14,5

TABLELANDS WITHIN THE LENGTH HISTOGRAMS

Common knowledge: Histograms of intron lengths have the peak occupying range of lengths ~50bp – 150bp.

New observation: Two tablelands to the right of the peak range can be found in the histograms of many species: the "close" (to the peak) and the "far" one. The histogram of A.carolinensis contains both tableland regions. The close one is ~160-240 bp (the close region is relatively short); the far one is ~500-1000 bp. Other species may have only one of the regions. E.g. birds have pronounced far region but very short (20 bp) close regions; it is hard to distinguish it from the far one. Human genome has pronounced short region (~150- ~200bp) but the far tableland (~400 bp - ~500bp) is significantly less pronounced than the far region in birds histograms. As to our knowledge the far tableland was reportd inly for D. rerio in [S. P. Moss et al. (2011) Genome Biol. Evol., 3:1187–1196] and explained by duplication events. The existence of close tablelands was independently observed by F. Kondrashov [personal communication].