# Statistical Analysis of Alignments of *Drosophila Melanogaster* Exons With Other *Drosophila* Species.

Tatiana V Astakhova[1], Vsevolod J. Makeev [2,3], Dmitry B. Malko [2,3] and Mikhail A. Roytberg[1]

[1]Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Puschino, Moscow Region, Russia; mroytberg@impb.psn.ru

[2]Institute of Genetics and Selection of Industrial Microorganisms, GosNIIGenetika, 117545 Moscow, Russia; makeev@genetica.ru

[3]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia; makeev@imb.ac.ru

The paper presents comparative study of constant exons from 12 complete genomes *Drosophila*; the *D. melanogaster* genome was used as a master one. It was shown that for all species level of similarity of internal exons is significantly greater than the levels of other exon classes (start-exons, stop-exons and one –exon genes). The relation between levels of similarity of other classes of exons depend on the phylogenetic distance between *D. melanogaster* and the specie under consideration.

## 1 Materials and Methods

The paper describes results of comparative study of constant exons from 12 complete genomes *Drosophila* (*D. melanogaster, D. simulans, D. sechellia, D. erecta, D. yakuba , D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. virilis, D. mojavensis, D. grimshawi*; constants exons are those presenting in all isoforms of a gene. The *D. melanogaster* genome was used as a master one. The chromosomes 4, 2L, 2R, 3L, 3R, X of the genome contain 13391 annotated genes. The genes contain 47331 *constant* exons. Using the programs BLAT [1] and Pro-Frame [2] we have aligned *D. melanogaster* genome against all other 11 genomes to find orthologs of each constant exon within each of 11 genomes.

The exons were divided into four classes (for each class we give the prefix that is used in tables): start exons (START), internal exons (INT), stop exons (STOP), one exon genes (ONE).

First, we have studied lengths of exons of different classes from different chromosomes (see Table 1). We would like to point out that average lengths of exons from chromosomes 2L and 3L greater than those of chromosomes 2R and 3R; the difference is most significant for internal and stop exons.

Second, we have studied similarity between exons and their orthologs. Average similarity SIM between exons of a given class (START, STOP, INT, ONE or ALL) of *D.Melanogaster* and their orthologs from a species can be calculated by formula SIM = T/L;. where

T = $\sum_{i=1,\ldots,N(CLASS)}$Length(DMel(i))*Identity(i, species);

Length(DMel(i)) is the length of i-th *D. Melanogaster* exon of the class;

Identity(i, species) is %id for alignment of the exon DMel(i) and its ortholog in the species; L is total length of all *D. Melanogaster* of the class having orthologs in the species.

## 2 Results and Discussion

The values of similarity SIM (in %%) for different species and classes are given in the Table 2.

The SIM value for internal exons is greater than the value for other classes of exons. The interrelations between SIM values other classes of exons depend on the phylogenetic distance of the species from *D.Melanogaster.* For the species from melanogaster subgroup (*D. simulans, D. sechellia, D. erecta, D. yakuba*) the highes SIM value show START exons, then STOP and ONE. For other species the situation is different: STOP is the best, then ONE, then START. Analogous data where obtained for each chromosome separately; the results for all chromosomes except chromosomes X and 4 are approximately the same. The SIM values for chromosomes X and 4 are significantly lower.

The biological reason of the greater conservativeness of internal exons compared to other classes of exons is unknown. A possible explanation may be connected with post-translation modification of proteins and thus greater importance of internal part of amino acid sequence. However this does not explain the lower level of similarity of one exon genes. This issue is the subject of further investigation.

**Proceedings of the 2nd International Conference on Mathematical Biology and Bioinformatics, Pushchino, Russia, 2008**

## References

[1] W.J..Kent. BLAT - The BLAST-Like Alignment Tool. Genome Research, vol. 12 (4), pages 656-664,, 2002

[2] A.A.Mironov, P.S.Novichkov, M.S.Gelfand . Pro-frame: similarity-based gene recognition in eukaryotic DNA sequences with errors. Bioinformatics, vol. 17 (1), , pages 13-15.. 2001

**Table 1.** Total number (a) and mean length (b) of exons from different classes and chromosomes.

**a)**

| | 2L | 3L | 2R | 3R | X | 4 |
|---|---|---|---|---|---|---|
| START | 1730 | 1858 | 1934 | 2390 | 1523 | 57 |
| INT | 4538 | 4498 | 5390 | 6492 | 3825 | 378 |
| STOP | 1801 | 1961 | 2030 | 2507 | 1593 | 63 |
| ONE | 613 | 564 | 517 | 599 | 463 | 7 |
| ALL | 8682 | 8881 | 9871 | 11988 | 7404 | 505 |

**b)**

| | 2L | 3L | 2R | 3R | X | 4 |
|---|---|---|---|---|---|---|
| START | 299.40 | 279.63 | 263.93 | 280.98 | 327.00 | 328.47 |
| INT | 396.35 | 371.88 | 343.34 | 358.14 | 423.20 | 401.58 |
| STOP | 479.00 | 501.54 | 437.85 | 445.28 | 517.45 | 526.03 |
| ONE | 834.26 | 1024.20 | 924.48 | 989.87 | 946.79 | 605.57 |
| ALL | 425.10 | 422.64 | 377.65 | 392.54 | 456.43 | 411.68 |

**Table 2.** Exon similarity measure SIM for different exon classes and different target species. The species are given in the incremental order by their phylogenetic distance from *D.Melanogaster*.

| | *ALL* | *START* | *INT* | *STOP* | *ONE* |
|---|---|---|---|---|---|
| D.mel. | 100 | 100 | 100 | 100 | 100 |
| D.sim | 92.228 | 92.763 | 93.005 | 91.601 | 90.134 |
| D.sec. | 94.351 | 93.942 | 95.724 | 93.405 | 91.653 |
| D.yak | 91.436 | 90.172 | 93.733 | 90.122 | 86.96 |
| D.ere. | 91.655 | 90.257 | 94.079 | 90.29 | 87.041 |
| D.ana | 80.964 | 75.37 | 85.092 | 78.76 | 74.601 |
| D.pse | 78.144 | 72.03 | 82.45 | 75.423 | 72.101 |
| D.per | 76.829 | 70.859 | 80.841 | 74.437 | 70.548 |
| D.wil | 75.239 | 68.354 | 79.61 | 72.017 | 69.356 |
| D.moj | 74.383 | 67.476 | 78.633 | 71.116 | 68.579 |
| D.vir | 74.507 | 66.857 | 78.949 | 71.304 | 68.922 |
| D.gri | 73.998 | 67.275 | 78.208 | 70.733 | 67.951 |