31. Addressing Drosophila Gene Duplications: Protein Sequence Redundancy and Genome Evolution

Carlos Quijano

CSIC-UAM Madrid

Gene duplication is a major force in genome remodelling during evolution. Published evidences shows that in Drosophila there is a considerable number of pairs of related genes that are physically linked. Our aim is to address globally their origin and functional fate by using bioinformatics approaches and experimental validation.

The Protein Universe is only a low volume projection in the protein sequence space, estimated in 5e+10 unique sequences. With 20 amino acids and an average length of 200 amino acids per sequence, the former contains 1,6e+260 possible proteins, more than the estimated number of electrons in the whole universe. On the other hand, it is clear that the distribution of observed protein families follows a power law, typical of scale free networks growing under preferential attachment. This means that the genetic material is mostly reutilized during evolution, with very few de novo generation of novelties. The birth process of genomic novelties is led by tandem duplications, whole or partial genome duplications, transposition and horizontal gene transfer. The common fate is degeneration and deletion, and somehow, a final distribution of sequences in stationary equilibrium and order. It is by the dynamic force introduced in the system by particular environmental interactions that the system evolves, as natural selection acts over populations of species with polymorphic genomes. Very little is known about how the entire system works, although some models have been proposed. We present here some bioinformatics algorithms developed by ourselves. Other current methods were not completely feasible to our aim of revealing how all the paralogous protein sequences are distributed. We selected Drosophila melanogaster as principal model organism due to the wide sources of experimental data available for further result analysis.

BLAST's HSPs Sum Statistics and Ordering as a new method for feasible and automated better multiple alignments:

Automated alignment of a complete proteome against itself using BLAST reveals only some of the most identical sequences (duplications). Other methods are not usable because their slow performance in exchange of little improvement. Ordering all the HSPs found by BLAST we can do an inference of the best feasible alignment, along all both sequences, using our algorithm implementation based on:

1 Powerfull Karlin & Altschul Sum Scores statistics
2 HSP order conservation
3 Buckhard Rost Cut off Equation for structural homology assumption

Family classification of related sequences without premises:

Current methods for classification of protein sequences produces non overlapping clusters with often unresolved phylogenies. For example, large groups of divergent Zinc Fingers and lonely sequences without any other familiar, when there are possible partners. Markovian techniques detects the better signal the more sequences while ignores divergent sequences when they are insufficient in number for obtaining statistical significance. We avoid both problems (unreticular relations and unresolution in clusters) using the previous algorithm, that gives BLAST a biological spirit, but respecting mathematical inference, and using our own clustering algorithm, that is based on:

1 A relations (duplications) matrix generated by iBLAST.
2 Intermediate Sequence Search (ISS) Strategy.
3 Cluster Growth ruled by Neighborly maximization.

Detection of Duplication Events between protein sequences and estimation of genomic linkages:

We have adressed all the genes that are family of all the individual genes in Drosophila melanogaster and every sequenced genome. We find that gene duplicates are linked in the genome with high significance. This signal is not due only to recent tandem duplications. When extracted from the database, high significance stands barely lowered ($p<0.001$, permutation test).