Poster C-16

Comparative analysis of nucleotide polymorphism levels across different taxonomic units.

Authors:
Natalia Petit (*Departamento de Genètica i Microbiología. Facultat de Ciències, Universitat Autònoma de Barcelona*)
Sonia Casillas (*Departamento de Genètica i Microbiología. Facultat de Ciències, Universitat Autònoma de Barcelona*)
Raquel Egea (*ica i Microbiología. Facultat de Ciències, Universitat Autònoma de Barcelona*)
Alfredo Ruiz (*ica i Microbiología. Facultat de Ciències, Universitat Autònoma de Barcelona*)
Antonio Barbadilla (*ica i Microbiología. Facultat de Ciències, Universitat Autònoma de Barcelona*)

Short Abstract: Two databases (DPDB and Mampol) compile sequences and estimate the levels of polymorphism of many organisms. Using this information, we designed and implemented a strategy to perform a global analysis of polymorphism data. The present study aims to describe and explain the polymorphism pattern in a wide range of species.

Long Abstract:
The analysis of the levels of polymorphism in different species shed light on the mechanisms of evolution of the genomes. Coding sequences of different genes and organisms have accumulated for years in Gene bank. At present, two databases (DPDB, Drosophila Polymorphism Data Base, http://dpdb.uab.es/dpdb 1 and Mampol, Mammalia polymorphism database, http://pda.uab.es/mampol) compile sequences and estimate the levels of polymorphism for 135 species of Drosophila and 3232 species of mammals. Using this information, we designed and implemented a strategy to perform a global analysis of polymorphism data. Since the evolution of coding sequences is associated with their function 2, 3, grouping genes according to their biological role (Gene Ontology 4) provide a way to achieve a global view of protein evolution. The present study aims to describe and explain the nucleotide polymorphism pattern in a wide range of species.
Methods: polymorphism data from 874 genes were obtained from DPDB 1.Genes were selected according to the number and length of the coding sequences used (more than six sequences with more than 100 bp or less than six sequences with more than 1000 bp), and the percentage of gaps less than 10 %. We ascertained the orthology between genes of different species by defining an identity threshold as follows: A random sample of 984 coding sequences was taken from the alignment of the seven species of Drosophila (Vista Genome Browser, http://pipeline.lbl.gov). For each coding sequence a Blastn was run on the gene annotated set (ref.) of D. melanogaster. The threshold value (<10-10) was estimated by averaging the e-values from each Blastn. By taking into account this parameter we found polymorphism data for 383 different genes belonging to 54 species of Drosophila. The dataset analyzed contain 26 species with polymorphism data for more than one gene, and 68 genes represented by more than one species.
Statistical analyses were performed to find out differences in the levels of synonymous (πs), non synonymous polymorphism (πn) and the ratio between them (πn/πs, which was considered an index of constraints), between genes,

organisms and functional categories of Gene Ontology. Mean differences were tested with a non parametric test (Kolmogorov-Smirnov), between each group of genes, organism or Gene Ontology category and a combination of all the remaining groups.
Gene Ontology categories of the genes at level 3, were obtained from Fatigo (http://fatigo.bioinfo.cipf.es 5) and the Gene Ontology (http://www.geneontology.org 4) web sites. The genes analyzed correspond to 49 categories of Biological Process, 45 categories of Molecular Function, and 20 categories of Cellular Component.

Results: This analysis indicated that 22 genes were different in their levels of synonymous polymorphism, but no significant differences were found in the levels of non synonymous polymorphism or constraints for any of the genes. On the contrary, non synonymous polymorphism was significantly smaller for D. miranda, D. santomea and D. sechellia in comparison to the other species, and constraints were only significantly different for D. yakuba and D. sechellia.
Among Biological Process categories with significantly less constraints (πn/πs more than the global mean) were those genes relating to behaviour and reproduction (GO: 0007610, GO: 0051705, GO: 0050795, GO: 0050876) and genes more constrained (πn/πs less than the global mean) were those concerning with segmentation, pattern specification and cell communication (GO: 0007389, GO: 0035282, GO: 0007154). For Molecular Function categories less constrained genes were those relating to protein, nucleotide and metal cluster binding (GO: 0005515, GO: 0000166, GO: 0051540) and the more constrained ones those grouped in ion and cofactor binding, hidrolase activity, isomerase activity, transcriptional repressor activity and oxidoreductase activity (GO: 0043167, GO: 0016787, GO: 0016853, GO:0016491 and GO:0016564). Lastly genes component of intracellular organelles, membrane and non-membrane bound organelles (GO: 0043229, GO: 0016020, GO: 0043228) seem be less constrained and genes component of membrane bound organelles and envelope of organelles (GO: 0043227, GO: 0031967) seem be more constrained.
The same analysis strategy will be implemented in the analysis of Mammalia polymorphism database (Mampol http://dpdb.uab.es/mampol) which contains nearly 1000 filtered polymorphic sets and the results will be compared to these in Drosophila.
References:
1. Casillas S, Petit N, Barbadilla A (2005) DPDB: a database for the storage, representation and analysis of polymorphism in the Drosophila genus. Bioinformatics 21: ii26-ii30.
2. Aris-Brosou S (2005) Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. Mol Biol Evol. 22:200-209.
3. Castillo-Davis CI, Kondrasov FA, Hartl D, Kulathinal RJ (2004). The functional genomic distribution in two animal phyla: Coevolution, genomic conflict, and constraint. Genome Res. 14:802-811.
4. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nat Genet. 25: 25-29.
5. Al-Shahrour, F., Minguez, P., Tárraga, J., Montaner, D., Alloza, E., Vaquerizas, J.MM., Conde, L., Blaschke, C., Vera, J. & Dopazo, J. (2006), BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Research, In Press.