Poster B-16

MamPol: Mammalia Polymorphism Database

Authors:
Raquel Egea (*Departament de Genetica i Microbiologia, Universitat Autonoma de Barcelona*)
Sonia Casillas (*Departament de Genetica i Microbiologia, Universitat Autonoma de Barcelona*)
Antonio Barbadilla (*Departament de Genetica i Microbiologia, Universitat Autonoma de Barcelona*)

Short Abstract: MamPol is a comprehensive database of nucleotide polymorphisms for the Mammalia class, including both the aligned sequences and their associated diversity estimates. The MamPol website integrates the information from the databases and offers several query interfaces, analysis tools and Help and Statistics sections.

Long Abstract:
Polymorphism studies are one of the main research areas of this genomic era. Recently, we have created DPDB1, a comprehensive secondary database which provides searchable collections of polymorphic sequences with their associated diversity measures for the Drosophila genus. We have improved and extended it to the Mammalia class in MamPol.

Methods: Data retrieving, calculation of the diversity measures and updating are done by PDA2, a pipeline made of a set of Perl modules that automates these processes. Using PDA we got all the well-annotated genomic DNA sequences available in Genbank for mammals, as well as their associated information and the cross-references to Popset. Sequences were grouped by species and gene name in 'polymorphic sets'. As this process is totally automated, sequences corresponding to the same gene but with differences in their names were assigned to different groups. To avoid this misclassification a list of synonymous gene names was manually created (available in the website).
For each polymorphic set, subgroups of homologous sequences were created corresponding to the different functional regions (genes, CDSs, exons, introns, UTRs and promoters) found in the sequence annotations.
Each subgroup was aligned with ClustalW3. The clustal parameters were optimized for alignments of polymorphic sequences. A 95% of similarity between each pair of sequences in the alignment was fixed as the minimum percentage score. If the score was lower, the sequence was extracted from the alignment. With this filter we can separate sequences that share erroneously the same gene name, and also sequences corresponding to different parts of the same gene.
In order to increase the quantity of information from the subgroups, different subsets were made taking into account the length of the sequences, implementing the estimation optimization algorithm developed by our group.
The final subsets of sequences were called the analysis units. Commonly used diversity measures were calculated on these analysis units, including polymorphism at synonymous and non-synonymous sites, linkage disequilibrium and codon bias.

Both the primary and secondary information were stored in relational MySQL databases. Sequences, polymorphic sets and analysis units were given a unique identification number to facilitate cross-database referencing and updating. The information is divided in three databases: (1) for primates, (2) for rodents, and (3) for the rest of mammals.
The databases are updated daily, looking for new sequences in Genbank and reanalyzing only the polymorphic sets affected.

The MamPol website: The MamPol website (http://pda.uab.es/mampol) integrates the information from the databases and offers several interfaces to query the database contents in different ways, provides analysis tools to reanalyze the polymorphic sets, a Help section where the user can find fully explanations about the web site, a Statistics section where the contents of the database are summarized and a series of links of interest classified in different categories.

1.-DATABASE QUERYING AND OUTPUT
The database contents can be queried through a web interface implemented as Perl CGI scripts based on SQL searches. The user can directly select the species of interest in the species list or a group in a higher taxonomic level in the taxonomic list. The later list is extensible and includes all the taxonomic levels from the mammalian class and allows selecting at any level. Gene names can also be selected in the genes list or in the gene name aliases list. In all these lists mitochondrial and nuclear data are separated, as well as data from rodents, primates and the rest of the mammals.

There are several query interfaces:
(1) The general search, which allows filtering for the diversity values and/or for the degree of confidences on the polymorphic set. The first output page lists all the polymorphic sets by organism, gene and analysis unit showing additional information about the quality of the alignment, the confidence on the data source and the date of the last update. A complete report for each analysis unit can be obtained through the corresponding link. It is also possible to reanalyze any polymorphic set with PDA. Furthermore, sequences can be directly downloaded in the fasta format.

(2) The graphical search allows the same filters and shows the distribution of any of the diversity parameters estimated. Each class has a link to view the corresponding polymorphic set as in the general search.

(3) The comparative search allows the user to select for any species or taxonomic group and to put filters for the quality of the alignments to compare the polymorphism between two or more groups. The output page gives the number of analysis units for each group and the means of the diversity values selected, except for Tajima’s D, which are divided into negatives, neutral and positives and the amount of each division is given. There is a link to the general search results page in each taxonomic group. Different functional regions are compared separately in order to avoid the duplication of the sequences.

2.- ANALYSIS TOOLS
The website includes a set of common analysis tools which work in our server, thus avoiding the necessity of connecting to other databases to reanalyse the data. These tools are ClustalW, Jalview, SNPs-Graphic and PDA.

3.- STATISTICS

The statistics section summarizes the contents of both the primary and secondary databases. It is updated daily, and includes tabular and graphic information. The information is divided in rodents, primates and the rest of the mammals data. And there is another subdivision between mitochondrial and nuclear data.

References:

1. Casillas, S., Petit, N. and Barbadilla, A. (2005) DPDB: a database for the storage, representation and anaylisis of polymorphism in the Drosophila genus. Bioinformatics 21: ii26-ii30.

2. Casillas, S. and Barbadilla, A. (2006) PDA v.2: improving the exploration and estimation of nucleotide polymorphism in large data sets of heterogeneous DNA. Nucleic Acid Res, In press.

3. Chenna, R. et al, (2003) Multilple sequence alignment with the Clustal series of programs. Nucleic Acid Res., 31, 3497-3500.