Investigating the gene order variability and non-coding sequences of metazoan mitochondrial genomes: design and construction of a mitogenomics database

Lupi R[1], D'Onorio De Meo P[2], Picardi E[3], D'Antonio M[2], Paoletti D[2], Castrignanò T[2], Pesole G[3,4], Gissi C[1]

Motivation

Most mitogenomics studies of Metazoa are focused on phylogenetic reconstructions, while there are only few comparative studies analysing structural features such as gene order and non-coding regions (NCR), or aimed at revealing the correlation between structural and functional mitochondrial (mt) features within an evolutionary framework. In order to stimulate comprehensive comparative analyses of under-investigated mtDNA features, such as gene order and NCRs, we have developed a new mtDNA database, named MitoZoa, whose main novelties are: (1) the improvement and curation of genome annotation, as a pre-requisite for database construction; (2) the definition and annotation of all NCRs, which can be easily retrieved based on their size and on the nature of flanking genes; (3) the definition of gene order as a string of standardized gene names and its storage in a FASTA-like format, easily manageable for gene order searches and downloadable for further analyses; (4) the possibility of selecting all mtDNAs of a given taxonomic group belonging to the same genus, which could be helpful in the investigation of mt evolutionary dynamics avoiding artefacts from saturation phenomena and to compare the overall mt evolutionary trend between different metazoan lineages.

Methods

The Mitozoa system consists of a relational database and a web interface: in particular, a set of PHP scripts allows the users, through the web interface, to build powerful queries without any knowledge of SQL. Data are output in web-table formats, moreover information and sequences can be downloaded in simple-text, excel or FASTA format. Mt annotations have been verified using a semi-automatic pipeline including several Python scripts. This pipeline focuses on the validation and correction of data on: molecule topology (circular or linear), exact denomination and strand location of tRNA and rRNA genes, boundaries of tRNA genes, non-canonical start codons in protein-coding genes, and partial status of the genome. As a general rule, all modifications introduced by the MitoZoa

[1] Dipartimento di Scienze Biomolecolari e Biotecnologie, Università degli Studi di Milano, Italy. [2] CASPUR, Italian Interuniversities Consortium for Supercomputing Applications, Roma, Italy [3] Dipartimento di Biochimica e Biologia Molecolare "E. Quagliariello", Università di Bari, Italy [4] Istituto Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Italy

reannotation process have been carried out in a conservative way, thus original gene annotations have been deleted only when certainly erroneous, while they have been retained and highlighted with “warning” notes in case of ambiguous validation. Mt entries have been enriched with relevant data lost in the primary entry (such as information on cryptic species and gender-specific mito-types) and also includes the standardization of mt gene names, which have been used as hidden aliases embedded in the database in order to: (1) write the whole mt gene order as a string of standardized gene names; (2) define the gene located upstream/downstream a NCR; (3) implement the single-step retrieval of all homologous genes belonging to a given taxonomic group.

Results

The MitoZoa database is freely available at http://www.caspur.it/mitozoa (Lupi et al. Mitochondrion 2010 10:192-9). Release 2.0 contains 2018 mtDNA sequences of Metazoa corresponding to: (1) genomes described as complete in the original entry; (2) partial mtDNAs with size & 7 kb, specifically selected as they are the only representatives of a metazoan phylum or members of congeneric pairs. Entries are shown in an EMBL-like format differing from the original EMBL format for the presence of four new fields and some new keys/qualifiers in the FT (standard EMBL fields have been used whenever possible, while new fields and sub-fields have been created only when necessary). Among novelties, the new FTkey “NCR” is used to annotate the non-coding regions of any size located between two consecutive genes (i.e. tRNA, rRNA or proteins). In addition, each NCR is associated with the new FTqualifier “code” which is related to species, NCR length, and NCR genomic position: this code has allowed implementing the selection of NCRs based on mtDNA position. The gene order is presented as a string of standardized gene names, with genes encoded by the reverse-complement strand preceded by a “minus” sign, and genes interrupted by introns or split in two parts indicated with the standardized gene name followed by the symbols “_5” or “_3” for the 5’- and 3’-end. The storage and download of gene order in a simple FASTA-like format facilitate downstream program applications and has not been intended for graphical representation. In addition, this format has simplified the implementation of a database search option able to find gene strings of any size within the entire database in few seconds. The MitoZoa database can be queried using a “General Search” menu alone or in combination with one of the following specialized menus: (1) Gene Order; (2) Non-Coding Region; and (3) Gene Content. Thus, using a specialized menu, the query can take advantage of all selection criteria available in the main “General Search” menu. Finally, MitoZoa site contains a page dedicated to major statistics on the type and number of errors corrected during the reannotation pipeline.

Contact e-mail

carmela.gissi@unimi.it