MitoZoa: a curated database of mitochondrial genomes of Metazoa specifically developed for genomics and phylogenetic analyses

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

[1] Dipartimento di Scienze Biomolecolari e Biotecnologie, Università 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 Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Bari, Italy

Motivation

Mitochondrial DNA (mtDNA) is the molecule of choice for phylogenetic reconstructions, especially among metazoans and its wide use is well-documented by the development of various specialized mt databases (mtDBs), such as GOBASE [1], OGRe [2], MitBASE [3], AMiGA [4], MamMiBase [5], Mitome [6] and the NCBI Organelle resources database [7]. These mtDBs are useful for retrieving sequences and basic information on all or some metazoan mtDNAs, even though many of them suffer from the same misannotations affecting the original RefSeq and EMBL entries, or from some flaws in the database structure and/or in the user interface, or implement the search for specific features in an ineffective way. For example, the NCBI Organelle Resources and GOBASE, which claim to be well-validated resources, still fail to rectify the large number of errors present in the original entries, while in OGRe, the noteworthy possibility to compare the mt gene order is allowed only via a graphic representation, thus its efficacy is questionable. Finally, in many mtDBs it is difficult to simultaneously extract a given mt feature for a large dataset. In conclusion, existing mtDBs are far from being powerful tools for thorough analyses of the copious mtDNA data publicly available. Here we describe a new specialized database, MitoZoa, collecting Metazoa mtDNA entries whose annotation has been enriched and drastically improved thanks to a semi-automatic pipeline.

Methods

The Mitozoa system consists of a relational database and a web interface. The data model is optimized for efficient storage and data retrieval. The used dbms is MySQL Enterprise 5.0 to ensure high availability of the system. A generalized parsing module simplifies the process of automatically incorporating and updating information. Using the SQL client we have the ability to perform a wide range of powerful queries. A set of PHP scripts allows the users, through the web interface to build queries without any knowledge of SQL. Data are output in web-table format, moreover information and sequences can be downloaded in excel and zip format, respectively. The first release of MitoZoa contains the >1200 metazoan mtDNA entries revised in Gissi et al [8], where it was reported that about 30% were affected by errors, and will be updated thanks to an automatic query via SRS@EBI. Entries are in the standard Embl format, and include a brand new Comment field listing all identified errors and introduced changes via standardized messages, along with information about the original RefSeq/EMBL entry. MitoZoa is maintained by a suite of scripts written in Python 2.5. Briefly, these scripts parse the information contained in the entries in order to standardize gene names, check for errors concerning the complete/partial status of the genome, and supply with notes or correct the original tRNA and rRNA gene annotations. In particular, the name of tRNA genes is completed by data on the anticodon, the recognized codons (in the case of more tRNAs loading the same amino acid), or other peculiarities. The identity and boundaries of tRNA genes are checked using tRNA-specific patterns defined and searched by PatSearch [9] and home-made scripts inspecting the tRNA length, respectively: if necessary, tRNAscan-SE [10] or Arwen [11] are applied to rectify or recalculate tRNA limits. A Python module extracts all intergenic non-coding regions (NCRs) and put them in the feature table (FT) as a new FTkey. Finally, the gene order is calculated and extracted into an easy-to-manage fasta-like format, where genes are reported with standardized names.

Results

MitoZoa represents a new and reliable tool for researches aiming to perform metazoan phylogenetic reconstructions or evolutionary genomics analyses of under-investigated mt features, such as NCRs and gene order. In fact, the database improves the annotation of the currently

available mtDNAs, thanks to an automatic rectification of known or foreseeable mistakes. High quality annotation standards are also guaranteed since the most troublesome entries are flagged for human intervention and a note is added when it is not possible to unambiguously resolve strange mt annotations. Besides standardizing gene names, MitoZoa is the first database focusing on the accuracy of tRNA genes, particularly their orientation, limits and recognized anticodon. The correctness of annotation is an important issue also for NCRs, which are inserted in each entry as a novel key of the FT, giving the user easy means for displaying/downloading them based on length and/or bordering genes. A similar consideration holds for gene order, which is easily downloadable in a fasta-like format apt to be input in programs analyzing gene order rearrangements. Finally, MitoZoa contains a page dedicated to major statistics on the type and number of errors corrected during the re-annotation pipeline, which will hopefully lead to better the quality of mtDNAs present in both specialized and primary databases.

References

(1) O'Brien, E. A., Zhang, Y., Wang, E., Marie, V., Badejoko, W., Lang, B. F. & Burger, G. (2009) GOBASE: an organelle genome database Nucleic Acids Res 37, D946-50.
(2) Jameson, D., Gibson, A. P., Hudelot, C. & Higgs, P. G. (2003) OGRe: a relational database for comparative analysis of mitochondrial genomes Nucleic Acids Res. 31, 202-6.
(3) Attimonelli, M., Altamura, N., Benne, R., Brennicke, A., Cooper, J. M., D'Elia, D., Montalvo, A., Pinto, B., De Robertis, M., Golik, P., Knoop, V., Lanave, C., Lazowska, J., Licciulli, F., Malladi, B. S., Memeo, F., Monnerot, M., Pasimeni, R., Pilbout, S., Schapira, A. H., Sloof, P. & Saccone, C. (2000) MitBASE : a comprehensive and integrated mitochondrial DNA database. The present status Nucleic Acids Res. 28, 148-52.
(4) Feijao, P. C., Neiva, L. S., de Azeredo-Espin, A. M. & Lessinger, A. C.(2006) AmiGA: the arthropodan mitochondrial genomes accessible database Bioinformatics.22,902-3.
(5) Vasconcelos, A. T., Guimaraes, A. C., Castelletti, C. H., Caruso, C. S., Ribeiro, C., Yokaichiya, F., Armoa, G. R., Pereira Gda, S., da Silva, I. T., Schrago, C. G., Fernandes, A. L., da Silveira, A. R., Carneiro, A. G., Carvalho, B. M., Viana, C. J., Gramkow, D., Lima, F. J., Correa, L. G., Mudado Mde, A., Nehab-Hess, P., Souza, R., Correa, R. L. & Russo, C. A. (2005) MamMiBase: a mitochondrial genome database for mammalian phylogenetic studies Bioinformatics.
21, 2566-7.
(6) Lee, Y. S., Oh, J., Kim, Y. U., Kim, N., Yang, S. & Hwang, U. W. (2008) Mitome: dynamic and interactive database for comparative mitochondrial genomics in metazoan animals Nucleic Acids Res. 36, D938-42.
(7) Wolfsberg, T. G., Schafer, S., Tatusov, R. L. & Tatusov, T. A. (2001) Organelle 66 of 144 18-02-2009 17:56 genome resource at NCBI Trends Biochem Sci. 26, 199-203.
(8) Gissi, C., Iannelli, F. & Pesole, G. (2008) Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species Heredity 101, 301-20.
(9) Pesole, G., Liuni, S. & D'Souza, M. (2000) PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance Bioinformatics. 16, 439-50.
(10) Lowe, T. M. & Eddy, S. R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence Nucleic Acids Res. 25, 955-64.
(11) Laslett, D. & Canback, B. (2008) ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences Bioinformatics. 24, 172-5.

Contact : carmela.gissi@unimi.it