MitoNuc: a database of nuclear genes encoding for mitochondrial proteins

D. Catalano[(1)], F. Licciulli[(1)], G. Grillo[(1)], S. Liuni[(1)], G. Pesole[(2)], C. Saccone[(1,3)] and D. D'Elia[(1)]

[(1)] Istituto di Tecnologie Biomediche - Sezione di Bari, CNR, Palazzo dei Geometri, Via Amendola 168/5, 70126 Bari - Italy
[(2)] Dipartimento di Fisiologia e Biochimica Generali, Università di Milano, Via Celoria, 26, 20133 Milano - Italy
[(3)] Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, Via Orabona, 4, 70126 Bari - Italy

Keywords. Bioinformatics, Database, Mitochondrion, Nuclear Genes for Mitochondrion

Introduction

Mitochondria are sub-cellular organelles, present in the majority of eukaryotic organisms, which play a central role in the energy metabolisms of cells. They are also involved in many other cellar processes such as apoptosis, aging and in a number of different human diseases, including Parkinson's, diabetes mellitus and Alzheimer's. Despite to their importance in the cell life maintenance, about the 95% of proteins, contributing to mitochondrial biogenesis and functional activities, are nuclear encoded, synthesized in the cytosol and targeted to mitochondria. The expression and assembling of these proteins are strictly dependent by the coordinated activities of the two genomes, mitochondrial and nuclear, but the molecular mechanisms and co-evolutionary processes of the cross-talk between these two genomes are still largely unknown. MitoNuc [1] is a specialized database of nuclear encoded mitochondrial proteins in Metazoa. It provides comprehensive data on genes and proteins consolidating information from external databases. These data include: gene sequence, structure and information from ENSEMBL [2], protein sequence and information from SWISSPROT [3], transcript sequence and structure from RefSeq [4] and UTRdb [5], disease information from OMIM [6]. Each database entry consists of a nuclear gene coding for a mitochondrial protein in a given species, and reports information on: species name and taxonomic classification; gene name, functional product, sub-cellular mitochondrial localization, protein tissue specificity, Enzyme Classification (EC) code for enzyme and disease data related to protein dysfunction. For each gene and gene product the Gene Ontology (GO) classification [7] with regard to molecular function, biological processes and cellular component is reported too. Links to external database resources are also provided. As far as the gene and transcript sequences data are concerned, in the previous MitoNuc releases they were extracted from the EMBL [8] related entries. Due to the high level of sequences redundancy in the primary database, the majority of MitoNuc entries contained more than one transcript and coding gene sequence for the same gene, thus introducing a remarkable redundancy level that affects the effectiveness of the database for sequence analysis aims. In order to remove redundancy we generated a MitoNuc section of gene and transcript sequences derived from those organisms whose genome sequence draft has been completed and annotated in ENSEMBL. These MitoNuc entries are available in the database section called "MitoNuc Genomics" that, at present, include the following species: *Homo sapiens, Rattus Norvegicus* and *Mus Musculus*. MitoNuc can be queried using the SRS Retrieval System [9] (http://www.ba.itb.cnr.it/srs/); the present release contains a total of 1344 entries among which 662 are collected in the MitoNuc Genomic section. The total number of species included in MitoNuc is about 64.

System and Methods

MitoNuc is a database conceptually structured in a relational schema and the Database Management System (DBMS) used for this application is MySQL. The database entries are generated from MySQL in a flat file format (EMBL format) to make it publicly available and retrievable through the SRS System Web interface and the EMBOSS package [10] utilities. Only the gene sequence is reported in the database, protein and transcript sequences can be dynamically extracted through the linking database function implemented in SRS from their external resources (SwissProt, RefSeq, UTRdb). Each database entry is annotated through a process automated thanks to the development of bio-perl scripts able to extract data from the external databases. Control procedures, including BLAST check [11] against protein sequences reported in the correlated public data resources (SWISSPROT and ENSEMBL) have been implemented, to avoid data inconsistencies. Protein sequences from all the different species present in MitoNuc are pair-wise aligned against the human protein sequence using the Needleman-Wunsch global alignment. Proteins whose sequence similarity is higher than the threshold fixed value of 60%, and fall into the same functional class, are multi-aligned using the CLUSTAL algorithm [12], manually controlled for consistence and grouped in Clusters. Each Cluster is named with the SWISSPROT Human identifier and groups homologous proteins from all the species present in MitoNuc. These data are collected in the database and will be soon correlated with the database entries so that they can be queried together and/or separately from entries data. As far as the query interface is concerned, the database can be queried combining different searching criteria and/or running multi-record queries. The multi-record querying can be run giving a list of reference values such as the entry identifiers (ID) from MitoNuc or external linked databases, the name of the genes or the GO ID, etc. The whole entries content or only part of them, such as gene and transcript sub-sequences (exons, introns, UTRs regions), can be extracted. Data can be saved locally in different file format using the SRS facilities.

Conclusion

MitoNuc database can be particularly useful for people interested in the study of the functional genomics and proteomics of nuclear gene and encoded proteins whose sub-cellular location is the mitochondrion. Indeed, in spite of the growing number of mitochondrial databases worldwide produced, the majority of them are species-specific, mainly human dedicated, or collect information on proteins only. Moreover, no one is able to allow the extraction of both, protein and gene sequences for a large set of these bio-molecules in only one time. Our goal is to make of MitoNuc the most reliable data resources until now produced in this field. To this aim we are already working on the improvement of its data content and on the enhancement of its Web interfaces and query support. Additional data that will be included are: SNPs information for the genes sequences, genomic intergenic region, functional protein domain information from PFAM [13]; protein structure data from PDB [14], protein family and family-specific signatures from InterPro [15]. The implementation of text mining procedures has also been planned for the retrieval and integration of protein-protein and DNA-protein interaction data available from external specialized data resources and for the integration of literature associated data from PUBMED. The database will be available in a table format too and statistics on its data set will be provided both in table and graphic format.

Acknowledgements

This project is funded within the grant of the PNR 2001-2003 (FIRB art.8) D.M. 199 - MIUR.

References

[1] Attimonelli, M. et al. MitoNuc: a database of nuclear genes coding for mitochondrial proteins. Update 2002. *Nucleic Acids Res.,* 30, 172-3, 2002.

[2] Birney E. et al. Ensembl 2004. *Nucleic Acids Res.*, 32, 468-70, 2004.

[3] Bairoch, A. and Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.,* 28, 45-48, 2000.

[4] Pruitt, K.D. and Maglott, D.R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, 29,137-40, 2001.

[5] Pesole, G. et al. UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res*., 30, 335-40,2002.

[6] Hamosh, A. et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, 30, 52-5, 2002.

[7] Harris, M.A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res,.* 32, 258-61, 2004.

[8] Kulikova, T. et al. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*., 32, 27-30, 2004.

[9] Zdobnov, E.M et al. The EBI SRS server – new features. *Bioinformatics Application Note*, 18, 1149-1150, 2002.

[10] Rice, P. et al. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics,* 16, 276-277, 2000.

[11] Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.,* 25, 3389-3402, 1997.

[12] Chenna, R. et al. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*., 31, 3497-500, 2003.

[13] Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res*., 32, 138-41, 2004.

[14] Kihara, D. and Skolnick, J. The PDB is a covering set of small protein structures. *J Mol Biol.*, 334,793-802, 2003.

[15] Mulder, N.J. et al. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*.,31, 315-8,2003.