322–324 Nucleic Acids Research, 2003, Vol. 31, No. 1
DOI: 10.1093/nar/gkg123

MitoDrome: a database of *Drosophila melanogaster* nuclear genes encoding proteins targeted to the mitochondrion

Marco Sardiello, Flavio Licciulli[2], Domenico Catalano[2], Marcella Attimonelli[1,*] and Corrado Caggese

Dipartimento di Genetica e Anatomia Patologica and [1]Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, Via E. Orabona 4, 70126 Bari, Italy and [2]Centro di Studio sui Mitocondri e Metabolismo Energetico CNR, Via Amendola 165/A, 70126 Bari, Italy

Received August 15, 2002; Accepted November 13, 2002

ABSTRACT

Mitochondria are organelles present in the cytoplasm of most eukaryotic cells; although they have their own DNA, the majority of the proteins necessary for a functional mitochondrion are coded by the nuclear DNA and only after transcription and translation they are imported in the mitochondrion as proteins. The primary role of the mitochondrion is electron transport and oxidative phosphorylation. Although it has been studied for a long time, the interest of researchers in mitochondria is still alive thanks to the discovery of mitochondrial role in apoptosis, aging and cancer. Aim of the MitoDrome database is to annotate the *Drosophila melanogaster* nuclear genes coding for mitochondrial proteins in order to contribute to the functional characterization of nuclear genes coding for mitochondrial proteins and to knowledge of gene diseases related to mitochondrial dysfunctions. Indeed *D. melanogaster* is one of the most studied organisms and a model for the Human genome. Data are derived from the comparison of Human mitochondrial proteins versus the *Drosophila* genome, ESTs and cDNA sequence data available in the FlyBase database. Links from the MitoDrome entries to the related homologous entries available in MitoNuC will be soon implemented. The MitoDrome database is available at http://bighost.area.ba.cnr.it/BIG/MitoDrome. Data are organised in a flat-file format and can be retrieved using the SRS system.

INTRODUCTION

Mitochondria are organelles present in the cytoplasm of most eukaryotic cells. Their primary role is electron transport and oxidative phosphorylation. All mitochondria studied to date have their own mitochondrial DNA and separate protein synthesising machinery. The majority of the proteins necessary for a mitochondrion to work are synthesised in the nucleus and then imported in mitochondria and thus only a small number of genes is localised in mitochondrial DNA. Therefore, the biogenesis of mitochondria is a complex mechanism requiring the contribution of both nuclear and mitochondrial genetic systems. Under this topic, we have already published the MitoNuC database (1) which annotates the Metazoa nuclear genes coding for mitochondrial functions; data in MitoNuC are derived from the SWISS-PROT Metazoa mitochondrial proteins further annotated with specialising information. The database MitoDrome, here presented, is aimed at contributing to the functional characterization of nuclear genes coding for mitochondrial proteins and to the knowledge of gene diseases related to mitochondrial dysfunctions. The 'Human disease catalogue' (2) lists about one hundred diseases related to mitochondrial dysfunction; moreover, the role of the mitochondrion in the generation of reactive oxygen species and in the activation of apoptosis has suggested the involvement of mitochondrial defects in a wide variety of degenerative diseases, in the process of aging and in cancer. However, many of the involved nuclear genes are still to be identified; an inventory and characterization of such genes as complete as possible is essential for a better understanding of molecular processes related to the mitochondrion. A contribution in this direction has been obtained by the group of genes published by Marc *et al.* (3) through the analysis of nuclear genes coding for mitochondrial proteins based on a microarray approach. Data in the database are derived from *in silico* analysis prevalently relying on Human versus *Drosophila melanogaster* (in the whole paper simply cited as *Drosophila*) biosequence comparisons. The comparison between DNA and protein sequences from different organisms is indeed a significant source of information on the function of eukaryotic genes and proteins involved in key biological processes, especially when coupled with *in vivo* studies using adequate animal models. *Drosophila* has a well-established role as a model organism for

*To whom correspondence should be addressed. Tel: +39 0805482100; Fax: +39 0805482607; Email: marcella@area.ba.cnr.it

genetic and molecular analysis and provides peculiar opportunities to answer biologically relevant questions. The main goal of our research is to generate a collection of *Drosophila* nuclear genes encoding mitochondrial proteins and to make them available in a specialized database. At present, the number of known *Drosophila* nuclear genes encoding mitochondrial proteins, whose function has been experimentally studied through genetic and molecular analyses, is relatively low, presumably because of the difficulty to isolate such genes in conventional genetic screens. However, the identification of most *Drosophila* nuclear genes encoding mitochondrial proteins has become feasible, due to the recently completed sequencing of the *Drosophila* genome and thanks to the availability, in public databases [GenBank (4) and FlyBase (5)], of an increasing number of *Drosophila* ESTs and complete cDNAs and of collections of strains where single transposable element insertions disrupt unique genes. The fruit fly, *Drosophila melanogaster*, database (FlyBase), is the official resource where data derived from literature and value-added data derived from *in silico* analysis are annotated. In particular the section GadFly (Genome Annotation Database in FlyBase) reports the annotations of data obtained through the application of gene prediction methods on the whole *Drosophila* genome.

Data generation

The strategy we adopt for the rapid localization in the genome of *Drosophila* nuclear genes encoding mitochondrial proteins and for their mutational analysis is based on searching the *Drosophila* public databases for similarities with known human mitochondrial proteins. This approach has been implemented according to the procedure described below:

(i) Select SWISS-PROT protein entries from *Homo sapiens* reporting in the description line the word 'mitochondrion' or 'mitochondrial'.

(ii) The resulting list contains 304 proteins in the SWISS-PROT release 40.35.

(iii) Each protein in the above list is analysed with TBLASTN against the *D. melanogaster* genome available at the Berkeley *Drosophila* Genome Project (BDGP) site.

(iv) The resulting highest scores are examined, so that, if the scores are significant enough, the resulting gene sequence and about 1000 nucleotides flanking the 5′ and 3′ gene ends are extracted from the scaffold available at the NCBI and stored in the MitoDrome database.

(v) The *Drosophila* gene sequence is analysed with BLASTN against the *D. melanogaster* ESTs database available on the BDGP and NCBI sites. Thanks to the comparison of the gene sequence against the ESTs, the gene architecture can be completely defined and reported in a map available through the database. In particular exon/intron boundaries and 5′ and 3′ flanking sequences are mapped.

(vi) The genomic sequence is also analysed in order to test through BDGP if a single transposable element P disrupts and/or misexpresses the *Drosophila* genes and to check in FlyBase for the presence of other stored mutant alleles. This is performed at the http://flybase.bio.indiana.edu/bin/fbidq.html?Fbti0004964&resultlist=fbti3992.data%5B site.

Data resulting from the above described procedure are stored in the MitoDrome database according to the data scheme reported below and available on the web site http://bighost.area.ba.cnr.it/BIG/MitoDrome.

Database organisation

Each gene that, according to the results of its homologous human protein, can be classified as '*in silico* predicted *Drosophila* mitochondrial gene' is annotated in a table available through the MitoDrome web site where the data are alphabetically ordered according to the protein name. What the table lists for each gene is: the MitoDrome Identifier hyper-linked to the complete MitoDrome entry in flat-file format, the gene name, the SWISS-PROT Human protein IDentifier, the protein name, the amino acidic length of both the Human- and *Drosophila*-predicted proteins, the percentage of identity and similarity resulting from the pair-wise alignment between Human- and *Drosophila*-predicted proteins, the *Drosophila* Gene Symbol as assigned in FlyBase, the corresponding FlyBase IDentifier, the cytogenetic localization on the genome as reported in the FlyBase entry and, if available, the P-element insertion symbol or its synonym as reported in FlyBase.

Data in the five latter columns are hyper-linked to:

- the pair-wise alignment. This can also be a multiple alignment if the same human protein produces more significant scores against different *Drosophila* proteins or if different human proteins report significant scores against a single *Drosophila* proteins. In this case, identity and similarity scores are more than one because they refer to each pair-wise comparison. The multiple alignment is obtained using the 'multalin' program (6) (http://prodes.toulouse.inra.fr/multalin/multalin.html);
- a text file containing
  - a Map describing the structure of the *Drosophila*-predicted gene, its mRNA sequence and its coding sequence architecture. By clicking on the map, the complete MitoDrome entry in flat-file format is displayed;
  - the mRNA sequence where the coding part is coloured in red;
  - the protein sequence;
- the related FlyBase entry;
- the FlyBase cytogenetic map of the chromosome containing the located gene;
- the P-element entry from FlyBase in case of positive hits of the *Drosophila*-predicted gene against the P-element database.

Data availability and retrieval

The database is available at the following address http://bighost.area.ba.cnr.it/BIG/MitoDrome. A link to the MitoDrome data table is implemented from the home page.

The whole database has been indexed on the CNR-Bari SRS server (http://bighost.area.ba.cnr.it/SRS) and is available through the SRS top page in the section 'mitochondrial databases'. Retrieval can be performed by gene or protein name, or by any signal and region annotated in the MitoDrome Feature tables.

Data content

The application of TBLASTN to the 304 human mitochondrial protein sequences selected from the SWISS-PROT database against the *Drosophila* Genome resulted in a high number of positive hits, thus allowing the production of 350 MitoDrome entries available through the MitoDrome table.

Future perspectives

The availability in MitoDrome of a significant number of genes, will facilitate a search for function-related conserved motifs in genes encoding mitochondrial proteins. Among the questions that may be addressed through such studies, there are the nature of the signals for polypeptide import into the mitochondrion or in specific compartments of the organelle, and of the signals involved in the coordination of the nuclear and mitochondrial genetic systems. The correlation of the genomic sequences with genetic mapping data will provide useful entry points for *in vivo* studies of gene function exploiting P-element insertions in the genes of interest. Such insertions present peculiar advantages for the genetic and molecular analysis, because revertants are readily obtainable by secondary mobilization of the element and frequent imprecise excision generates useful new mutant alleles. Finally the availability of the *Anopheles gambiae* genome (http://www.ncbi.nlm.nih.gov/cgibin/Entrez/map_search?chr=agambiae.inf) will allow us to analyse it by applying the same procedure presented here for the *D. melanogaster* genome and to annotate the resulting data in a MitoDrome twin database.

ACKNOWLEDGEMENTS

This work has been supported by 'Ministero Università e Ricerca Scientifica', Italy (PRIN99, Programma Biotecnologie legge 95/95-MURST 5%; Progetto MURST Cluster C03/2000, CEGBA).

REFERENCES

1. Attimonelli,M., Catalano, D., Gissi,C., Grillo,G., Licciulli,F., Liuni,S., Santamaria,M., Pesole,G. and Saccone,C. (2002) MitoNuc: a database of nuclear genes coding for mitochondrial proteins. *Nucleic Acids Res.*, **30**, 172–173.
2. Fortini,M.E., Skupski,M.P., Boguski,M.S. and Hariharan,I.K. (2000) A survey of Human disease gene counterparts in the *Drosophila* genome. *J. Cell Biol.*, **150**, F23–F29.
3. Marc,P., Margeot,A., Devaux,F., Blugeon,C., Corral-Debrinski,M. and Jacq,C. (2002) Genome-wide analysis of mRNAs targeted to yeast mitochondria. *EMBO Rep.*, **3**, 159–164.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
5. The FlyBase Consortium (2002) The FlyBase database of the *Drosophila* genome projects and community literature, *Nucleic Acids Res.*, **30**, 106–108.
6. Corpet,F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.*, **16**, 10881–10890.