Nucleic Acids Research, 1999, Vol. 27, No. 1 **143–146**

Update of the Human MitBASE database

M. Attimonelli[1,*], J. M. Cooper[2], D. D'Elia[3], A. de Montalvo[4], M. De Robertis[5], H. Lehväslaiho[6], S. B. Malladi[6], F. Memeo[1], K. Stevens[2,7], A. H. V. Schapira[2,7] and C. Saccone[1]

[1]Dipartimento di Biochimica e Biologia Molecolare, Università degli Studi di Bari, 70126 Bari, Italy, [2]University Department of Clinical Neurosciences, Royal Free and University College Medical School, UCL, London, UK, [3]Area di Ricerca di Bari, CNR, 70126 Bari, Italy, [4]Departamento de Biología Molecular, Hospital Universitario 'Marqués de Valdecilla', Universidad de Cantabria, 39011 Santander, Spain, [5]Centro di Studio sui Mitocondri e Metabolismo Energetico, CNR, 70126 Bari, Italy, [6]EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton Hall, Hinxton, Cambridge, UK and [7]University Department of Clinical Neurology, Institute of Neurology, London, UK

Received October 5, 1998; Accepted October 8, 1998

ABSTRACT

Human MitBASE is a database collecting human mtDNA variants. This database is part of a greater mitochondrial genome database (MitBASE) funded within the EU Biotech Program. The present paper reports the recent improvements in data structure, data quality and data quantity. As far as the database structure is concerned it is now fully designed and implemented. Based on the previously described structure some changes have been made to optimise both data input and data quality. Cross-references with other bio-databases (EMBL, OMIM, MEDLINE) have been implemented. Human MitBASE data can be queried with the MitBASE Simple Query System (http://www.ebi.ac.uk/htbin/Mitbase/mitbase.pl) and with SRS at the EBI under the 'Mutation' section (http://srs.ebi.ac.uk/srs5/). At present the Human MitBASE node contains ~5000 variants related to studies investigating population polymorphisms and pathologies.

INTRODUCTION

Since the observation of mitochondrial DNA (mtDNA) mutations in patients with chronic progressive external ophthalmoplegia or Leber's hereditary optic neuropathy in 1998 (1,2), a growing number of mtDNA mutations have been associated with a wide variety of diseases (3). In addition to the patients with the 'classical' mitochondrial diseases mtDNA mutations have been found in patients with diabetes and heart failure (4–6). Moreover, mtDNA polymorphisms have been associated with Parkinson's and Alzheimer diseases (7,8) and age-related accumulation of somatic mtDNA mutations in post-mitotic tissues have been related to the aging process (9–11). Finally, due to its relatively high evolutionary rate mtDNA represents one of the major tools available for evolutionary studies of populations (12,13). Indeed the D-loop containing region of human mtDNA has a very high nucleotide substitution rate in two peripheral domains, the hypervariable regions I and II (HV-I and HV-II), a characteristic used extensively to study the origin of modern man.

The quantity of molecular and clinical data available from research groups interested in mitochondrial diseases and the number of variant sequences collected from population studies are growing exponentially; however, it would be very difficult to carry out statistical analyses and to obtain trustworthy results without the bioinformatic support.

The primary focus of the Human MitBASE database is to collect all the data available worldwide relevant to mitochondrial diseases and to mitochondrial DNA intraspecies diversity in a relational database and to set up an ad hoc query system to allow the retrieval of all this information.

Human mtDNA D-loop data previously structured by our group in MmtDB (14) (http://WWW.ba.cnr.it/~areamt08/MmtDBWWW.htm) have been completely stored in the Human MitBASE node of MitBASE database (15).

In MitBASE the complete human dataset has been defined distinguishing molecular from clinical and pathological data. The molecular dataset is under the responsibility of the Bari MitBASE group, while the clinical and pathological dataset is under the responsibility of the London MitBASE group.

Data are locally managed through Microsoft Access and then centrally stored at the EBI into the ORACLE MitBASE database (15).

DATA SOURCES

Human data are retrieved from bibliographic (MEDLINE) and from primary (EMBL data library and GenBank) databases (16,17). After careful revision these data are stored in MitBASE. Congress proceedings and unpublished data kindly provided by the authors are also included. The human data are coded using as

*To whom correspondence should be addressed. Tel: +39 080 548 2130; Fax: +39 080 548 4467; Email: marcella@area.ba.cnr.it

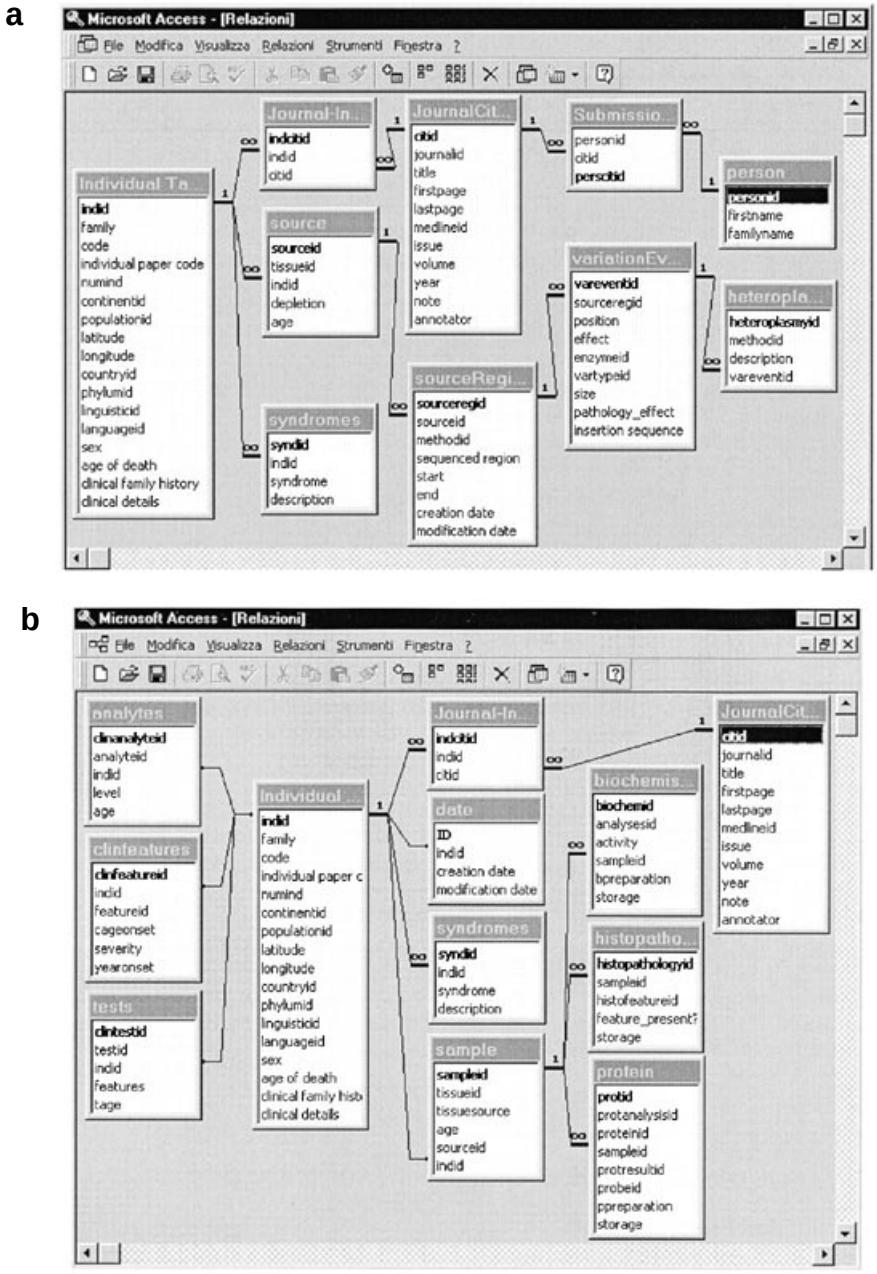

Figure 1. The Microsoft Access Human MitBASE structure: (**a**) the molecular structure and (**b**) the clinical/pathological structure. Data tables and links among them are shown. The fields reported in bold are the identifiers associated to the information in each table automatically updated during data input. The name of the tables describes their content.

a reference the nucleotide sequence published by Anderson *et al.* in 1981 (18). Despite being a hybrid derived from placental mtDNA and in part from HeLa cell mtDNA Anderson's sequence represents an important reference for human studies.

HUMAN MitBASE DATA STRUCTURE

The general structuring of the human dataset was already described in the previous paper (19). Each entry in the Human MitBASE database is related to a variant, which is defined as a mtDNA fragment with a different pattern of variation events with respect to the reference sequence.

The Human MitBASE data structure (Fig. 1) is composed of two sub-structures related to molecular (Fig. 1a) and clinical/pathological data (Fig. 1b).

Data structure is organised into data tables (reported in Fig. 1) and control value tables 'CV_' (tables not reported). Data tables contain specific data connected to each variant, whereas control value tables contain lists of values generally applicable to any variant. Links between data tables and control value tables have been defined in order to ensure a comprehensive integration of the available information.

All the tables in the structure are linked through the Individual table, which allows a targeted query to be built. This table

correlates the molecular sub-structure with the clinical/pathological sub-structure supported by the tables relevant to bibliographic information (JournalCitation, JournalInd and CV_Journal).

Each entry in the database is related to a nucleotide sequence variant located in a specific region of the mtDNA extracted from a tissue of an individual. This entry is identified by a sourceregion_id automatically generated. In this way the same individual can be associated to more than one entry if different regions of its genome have been analysed or if the same region has been analysed in different tissues. Each individual is identified in MitBASE by a pedigree code and/or an individual paper code. The pedigree code is composed of a family code, which identifies the single individual or a family when a pedigree study is reported, followed by characters, Roman and Arabic numbers, to mark the individual in the generation. An internal cross-referencing implemented in the ORACLE database will allow all the entries relevant to the same individual to be collected.

The information annotated in the molecular human MitBASE sub-structure include: analysed mtDNA region, experimental method used for the analysis, tissue or cell lines used for the molecular studies, sex, age and population data of the subject and information about his/her geographical and linguistic origin. Information about the type of variation occurred (substitution, deletion, insertion), the variation location, restriction site gain or loss are also reported.

In the 'clinical sub-structure' the following sub-groups have been defined based on different types of analyses carried out on the patients: clinical, histopathological, analyte and biochemistry features.

FLATFILE FORMAT STRUCTURE

A flatfile (ff) format has been fully designed and implemented as output of the Human MitBASE Simple Query System (http://www.ebi.ac.uk/htbin/Mitbase/mitbase.pl). This query system allows searching of human variants by gene name.

The ff format follows the rules agreed with the other partners of the MitBASE project (15). It reports information common to all MitBASE nodes (entry date, taxonomy, bibliography and cross-referencing line for the link with other biological databases) followed by the specific human MitBASE data. A general scheme of the present flatfile format is shown in Figure 2.

Individual, clinical, histopathological, biochemical and pathological data are codified adopting new two letter codes based on the model of the EMBL datalibrary 'FT' lines: there are feature key, feature qualifiers and feature description. The major benefit of this ff structuring is its implementation in SRS.

SEARCHING WITH SRS

In collaboration with the EBI HmutDB project (20) the Human MitBASE database has been incorporated into the EBI SRS server under the 'Mutations' section (http://srs.ebi.ac.uk/). All the fields, excluding the sequence, are indexed allowing detailed queries that can be combined with arbitrary complexity.

Desired fields can be selected and written out for further statistical analysis.

ACKNOWLEDGEMENT

This work has been funded under the EU Biotechnology programme, contract number: BIO4 CT950160.

```
ID      Entry name (based on sourceregid)  -  Variant sequence length
DE      Description of the variant
DT      Entry date
DT      Modification date
SM      Reference sequence entryname
OS      Organism name
OC      Organism Classification based on NCBI Taxonomy
RN      Bibliographic Reference numbering
RX      Medline code
RA      Bibliographic reference authors names
RT      Bibliographic reference title
RL      Bibliographic reference
IB      Individual block numbering
IN      Number of individuals
AG      Age of death
SX      Sex
IS      Clinical details
PD      Pedigree code
CC      Continent
CO      Country
CP      Population group
CY      Phylum
CL      Linguistic group
CD      Dialect
SY      Syndrome Achronym
SY              /note='syndrome description'
CF      Clinical Features
CF              /age_onset='value'
CF              /severity='value'
CF              /year_onset='value'
CF              /family_history='value'
CT      Clinical Test
CT              /clinical_test_feat='value'
CT              /test_age='value'
AN      Analytes
AN              /level='value'
AN              /age='value'
BN      Biochem-Hist-Sample numbering
BS      Tissueid description
BS              /source_age='value'
BS              /tissue_source='value'
BC      Biochem analyses
BC              /level='value'
BC              /storagage='value'
BC              /preparation='value'
PR      Protein analysis
PR              /protein_analysed='value'
PR              /protein_storage='value'
PR              /protein_analysis_result='value'
PR              /probe='value'
PR              /protein_preparation='value'
HS      Histopathologycal feature
HS              /storage='value'
HS              /feature_present='value'
SN      Molecular Sample numbering
SO      Molecular Sample tissue or cell lines
SO              /sample_age='value'
SO              /note='depletion field'
CC      General Comments
DR      Cross-referencing  to other databases
FH      Key         Location/Qualifiers
FH
FT      variation   Location of variation
FT              /gene=' gene_name'
FT              /sequence_class='value'
FT              /substitution='variation event value'
FT              /complement='complemented variation event value'*
FT              /deletion
FT              /deletion='single nucleotide deletion value'
FT              /insertion='nucleotide insertion value'
FT              /aa_change='aa change value'
FT              /note='enzyme effect'
FT              /note='conflict' or 'pathology associated'
FT              /heteroplasmy='value'
FT              /heteroplasmy_method='value'
EE      Experimental Method
AR      Analysed Region
SQ
Nucleotide sequence....
//

* if the gene is on the Complemetary strand
```

Figure 2. Flatfile format of Human MitBASE database.

REFERENCES

1 Holt,I.J., Harding,A.E. and Morgan-Hughes,J.A. (1988) *Nature*, **331**, 717–719.
2 Wallace,D.C., Singh,G., Lott,M.T., Hodge,J.A., Schurr,T.G., Lezza,A.M., Elsa,L.J.,II, Nikoskelainen,E.K. *et al.* (1988) *Science*, **242**, 1427–1430.
3 Suomalainen,A. (1997) *Ann. Med*., **29**, 235–246.
4 Paquis-Flucklinger,V., Vialettes,B., Canivet,B., Freychet,P., Hieronimus,S., Vague,P., Saunieres,A. and Desnuelle,C. (1997) *J. Annu. Diabetol Hotel Dieu*., 25–31.
5 Ozawa,T. (1995) *Eur Heart J*., **16**, 10–14.
6 Yoshida,R., Ishida,Y., Hozumi,T., Ueno,H., Kishimoto,M., Kasuga,M. and Kazumi,T. (1994) *Lancet*, **344**, 1375.
7 Shoffner,J.M., Brown,M.D., Torroni,A., Lott,M.T., Cabell,M.F., Mirra,S.S., Beal,M.F., Yang,C.C., Gearing,M., Salvo,R., *et al.* (1993) *Genomics*, **17**, 171–184.
8 Brown,M.D., Shoffner,J., Kim,Y.L., Jun,A.S., Graham,B.H., Cabell,M.F., Gurley,D.S. and Wallace,D.C. (1996) *Am. J. Med. Genet*., **61**, 283–289.
9 Ames,B.N., Shigenaga,M.K. and Hagenm,T.M. (1995) *Biochim. Biophys. Acta*, **1271**, 165–170.
10 Miquel,J. (1998) *Exp. Gerontol*., **33**, 113–126.
11 Wallace,D.C., Brown,M.D., Melov,S., Graham,B. and Lott,M. (1998) *Biofactors*, **7**, 187–190.
12 Saccone,C. (1994) *Curr. Opin. Genet. Dev*., **4**, 875–881.
13 Stoneking,M. and Soodyall,H. (1996) *Curr. Opin. Genet. Dev*., **6**, 731–736.
14 Attimonelli,M., Calò,D., De Montalvo,A., Lanave,C., Sasanelli,D., Tommaseo Ponzetta,M. and Saccone,C. (1998) *Nucleic Acids Res*., **26**, 120–125.
15 Attimonelli,M., Altamura,N., Benne,R., Boyen,C., Brennicke,A., Carone,A., Cooper,J.M., D'Elia,D., De Montalvo,A., de Pinto,B., De Robertis,M., Golik,P., Grienenberger,J.M., Knoop,V., Lanave,C., Lazowska,J., Lemagnen,A., Malladi,B.S., Memeo,F., Monnerot,M., Pilbout,S., Schapira,A.H.V., Sloof,P., Slonimski,P., Stevens,K. and Saccone,C. (1999) *Nucleic Acids Res*., **27**, 128–133.
16 Stoesser,G., Sterk,P., Tuli,M.A., Stoehr,P.J. and Cameron,G.N. (1998) *Nucleic Acids Res*., **26**, 8–15.
17 Benson,D.A., Boguski,M.S., Lipman,D.J. and Ostell,J. (1998) *Nucleic Acids Res*., **26**, 1–7.
18 Anderson,S., Bankier,A.T., Barrell,B.G., Debrujin,M.H., Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A., Sanger,F., *et al*., (1981) *Nature*, **290**, 457–465.
19 Attimonelli,M., Calò,D., Cooper,J.M., de Montalvo,A., Licciulli,F., Sasanelli,D., Stevens,K., Malladi,B.S., Saccone,C. and Schapira,A.H.V. (1998) *Nucleic Acids Res*., **26**, 116–119.
20 Lehväslaiho,H., Ashburner,M. and Etzold,T. (1998) *Trends Genet*.,**14**, 205–206.