Vertebrate MitBASE: a specialised database on vertebrate mitochondrial DNA sequences

A. Carone*, S. B. Malladi1, M. Attimonelli and C. Saccone

Dipartimento di Biochimica e Biologia Molecolare, Universita’ degli studi di Bari. Via E. Orabona 4, 70126 Bari, Italy and 1EBI, Hinxton Hall, Hinxton, Cambridge, UK

Received September 11, 1998; Revised October 15, 1998; Accepted October 27, 1998

ABSTRACT

Vertebrate MitBASE is a specialized database where all the vertebrate mitochondrial DNA entries from primary databases are collected, revised and integrated with new information emerging from the literature. Variant sequences are also analyzed, aligned and linked to reference sequences. Data related to the same species and fragment can be viewed over the WWW. The database has a flexible interface and a retrieval system to help non-expert users and contains information not currently available in the primary databases. Vertebrate MitBASE is now available through the MitBASE home page at URL: http://www.ebi.ac.uk/htbin/Mitbase/mitbase.pl. This work is part of a larger project, MitBASE which is a network of databases covering the full panorama of knowledge on mitochondrial DNA from protists to human sequences.

INTRODUCTION

During the last few years the importance of bioinformatics for the scientific community has grown following the increased number of biosequences available. A few years ago the EMBL (1) and GenBank (2) databases were highly representative of the amount of nucleic acids sequences available at that time for data-retrieval and queries. The enormous amount of data which is now submitted daily to the primary databases is due to the bench technical support given mainly by PCR (Polymerase Chain Reaction) which enables the fast sequencing of DNA fragments, genes and genomes and also to the easy use of network resources. The growth in database entries during the last decade and the parallel increasing amount of redundancies and mistakes has produced a change in database production.

These simple considerations lead to a question: how can specific data be easily retrieved from this ocean of information and moreover how can retrieved information be free from errors and really representative of the data available in the primary data-banks? An expert in bioinformatics has an array of appropriate tools for answering these questions but is it the same for a bench-scientist?

A possible solution can be found in the creation of a specialized database containing information related to a limited set of data. Following this idea, recently a large number of specialized databases whose data content is restricted to a well-defined subject and field, e.g., proteins, SWISS-PROT (3), mitochondrial and human mitochondrial DNA data GOBASE (4), MmtDB (5), MITOMAP (6), yeast sequences,YPD (7) and many more have been created.

Here we present a mitochondrial DNA vertebrate database not inclusive of the human data. This work is part of a more extensive EU-funded project, MitBASE (manuscript published Nucleic Acids Research Database Issue 1999), (EU BIO4-CT950160) co-ordinated by Prof. C. Saccone, structured in the ORACLE database management system and resident at the EBI in Hinxton, UK and now available on the Web at URL: http://www.ebi.ac.uk/htbin/Mitbase/mitbase.pl

DATA SOURCES

Data are mainly derived from the primary databases, EMBL data library and GenBank. The data from these main sources have been enhanced through the addition of information from the available literature and personal communications.

After the optimisation of queries over the full range of the primary databases by using different retrieval systems, SRS (8), ACNUC (9) and Entrez (10), with the purpose of cross-matching among retrieved data, we have standardized a set of queries

Table 1. Data source

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Number of EMBL entries (Release 54, 1998)</th>
<th>Number of entries present in MitBASE</th>
<th>Number of bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fish</td>
<td>2587</td>
<td>946</td>
<td>472513</td>
</tr>
<tr>
<td>Amphibians</td>
<td>508</td>
<td>151</td>
<td>90160</td>
</tr>
<tr>
<td>Reptiles</td>
<td>1242</td>
<td>375</td>
<td>178140</td>
</tr>
<tr>
<td>Aves</td>
<td>1800</td>
<td>538</td>
<td>385796</td>
</tr>
<tr>
<td>Mammals</td>
<td>4563</td>
<td>722</td>
<td>764789</td>
</tr>
<tr>
<td>Chordata</td>
<td>21</td>
<td></td>
<td>1891368</td>
</tr>
<tr>
<td>Total</td>
<td>10721</td>
<td>2732</td>
<td></td>
</tr>
</tbody>
</table>

Number of mitochondrial DNA vertebrate entries available in EMBL data library at release 54 (1998), subdivided in taxa. The first column represents raw data which have not been proofread for redundancies and errors while the second shows the corrected data already available in our database. The number of bases corresponding to the entries are also reported.

*To whom correspondence should be addressed. Tel: +39 080 548 2180; Fax: +39 080 548 4467; Email: areasc15@area.ba.cnr.it
covering the full vertebrate sequences subdivided in taxa: Fish, Aves, Reptiles, Amphibians, Mammals and Chordata (not inclusive of the previous five groups). Table 1 summarises the vertebrate data available in MitBASE at present.

**DATA STRUCTURE**

Data have been organised into two well-defined groups:

(i) citations, cross-references to other databases and nucleotide sequences. This information is already resident in the EMBL data library and only have to be checked, properly revised and often integrated or modified;

(ii) value-added information derived from publications or personal communications.

Our purpose has been the integration of information in the primary databases with details on bench techniques such as PCR details or sequencing and cloning methods, population localization, identification and classification of conservative sequences blocks (e.g., TAS, ETAS, CSB) (11), computational methods and programs applied to the sequences described in the entries.

To avoid time-consuming procedures such as re-entering existing data from group 1 we have designed a Web interface developed at the EBI within the MitBASE project and available under password from the MitBASE home page through the option ‘Submission of Data’. The interface enables the entries to be loaded from the EMBL data library into an ‘editable’ form. Entries can be selected by entry name or accession number. After choosing from a pull-down menu the fields of interest, the program outputs to the screen the strings of characters related to the desired fields in manageable blocks ready for editing. After the data are checked and revised this information is parsed and loaded into the centralized MitBASE database structured under the ORACLE management system.

Data related to the second group are locally stored in an MS Access database. It shows a single form with all the necessary fields linked to a complete set of lookup tables containing repetitive strings of characters to enable fast data entry. Pop-down menus are also available, minimizing syntax errors. This information, periodically released at the EBI, is parsed and loaded into the ORACLE database. Table 2 gives a schematic representation of the data-set structure containing data from both groups 1 and 2.

**TREATMENT OF VARIANT SEQUENCES**

When the content of the selected vertebrate data is explored, it becomes evident that there is a large numbers of sequences (Table 3) related to the same organism and the same gene or fragment. All these entries can be defined as variants of a sequence where the reference sequence can be a real fragment with a biological meaning or a synthetic sequence created by combining sequences. Figure 1 shows the possible cases of reference sequence creation. The simpler case (Fig. 1a) is the presence of a complete mitochondrial genome that covers all the possible variants. At present we only have 48 complete mitochondrial genomes available on the Web and over 1500 species with partial sequence data. This gives a clear overview of the biased trend in
this research area which concentrates on restricted number of species (e.g., *Homo sapiens* or *Rattus norvegicus*) and genes (e.g., ribosomal genes and cytochrome b gene). Therefore most of the sequences may not be referred to a complete genome but must be referred to fragments containing a part of a gene or several genes (Fig. 1b).

The treating of the variant sequences is one of the main goals of this project where users could retrieve a complete set of entries related to variants within a single organism where all the mutation details will be clearly described.

The retrieval of the variant sequence data has been performed with the program VARCLU (S.B. Malladi, manuscript in preparation) written in C language and running under a Windows environment. VARCLU can process a Clustal V or W (12) format file giving as output a text format file containing point mutations, insertions and deletions compared to a reference sequence included in the alignment. We could therefore codify most of the variants keeping off the more complex cases represented by overlapping fragments. We are now producing a new program able to automatically generate a consensus sequence from a set of fragments. This program will work coupled with VARCLU allowing the detection of all the possible variants and reference sequences.

**DATA-RETRIEVAL**

Vertebrate data can be retrieved using the ‘Simple Query System’ option available from the MitBASE home page. The vertebrate Simple Query System is part of a more comprehensive tool resident in the MitBASE home page. This retrieval system can query the vertebrate database following two pathways: (i) by species or (ii) by gene. Once the vertebrate section and the path are chosen, data can be displayed in different formats: (a) a flat file version or (b) a descriptive format. The first one is based on the most common EMBL flat file standard but contains a specific organisation of the identifiers common to all MitBASE outputs plus several new identifiers specially made for the requirements of vertebrate flat files. This output can be easily exported for data analyses and will be soon implemented in SRS (8). A clickable cross-referencing is also being set-up. When the descriptive output format is chosen, the same information is displayed identified by full descriptive strings of characters. This format has been proposed to encourage the use of databases by non-expert users who can now enjoy a descriptive and user-friendly output.

**ACKNOWLEDGEMENTS**

Thanks to Dr Matteo di Tommaso for the informatics support to the design of the database and to Dr D. G. Bembo for his suggestions during the manuscript preparation. This work has been funded under the EU Biotechnology programme, contract number: BIO4 CT950160.

**REFERENCES**