Databases and ontologies

AMiGA: the arthropodan mitochondrial genomes accessible database

Pedro Cipriano Feijão, Lissiene Silva Neiva, Ana Maria Lima de Azeredo-Espin and Ana Cláudia Lessinger*

Laboratório de Genética Animal, Centro de Biologia Molecular e Engenharia Genética (CBMEG), Universidade Estadual de Campinas (UNICAMP), CP 6010, CEP 13035-875, Campinas, São Paulo, Brazil

Received on September 28, 2005; revised on January 20, 2006; accepted on January 23, 2006
Advance Access publication January 29, 2006
Associate Editor: Alfonso Valencia

ABSTRACT

Summary: The Arthropod Mitochondrial Genomes Accessible database (AMiGA) is a relational database developed to help in managing access to the increasing amount of data arising from developments in arthropod mitochondrial genomics (136 mitochondrial genomes as of September 2005). The strengths of AMiGA include (1) a more accessible and up-to-date database containing a more comprehensive set of mitochondrial genomes for this phylum, (2) the provision of flexible search options for retrieving detailed information such as bibliographical data, genomic graphics, FASTA sequences and taxonomical status, (3) the possibility of enhanced comparative analyses by multiple alignment of single or concatenated sets of genes, (4) more accurate and updated information resulting from a specific curation process called AMiGA Notes and (5) the possibility of including unpublished sequences in a password-restricted area for comparative analysis with the other sequences stored in the database.

Availability: http://amiga.cbmeg.unicamp.br
Contact: lessinger@amiga.cbmeg.unicamp.br
Supplementary information: Detailed information, including an illustrated tutorial, is available from the above URL.

1 INTRODUCTION

Because of characteristics such as a reduced size, haploidy and a conserved gene content (37 genes in most cases: 13 protein-coding genes, 22 tRNAs and 2 rRNAs), the animal mitochondrial genome has been widely used as an important source of data for evolutionary studies (Boore, 1999). Primary sequence databases such as GenBank (Benson et al., 2005), maintained by the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov), the European Molecular Biology Laboratory (EMBL) (http://www.ebi.ac.uk), and the DNA Databank of Japan (DDBJ) (http://www.ddbj.nig.ac.jp/), have drastically improved and stimulated evolution studies based on comparative analyses of DNA sequences. More exclusive databases, such as Gobase (O’Brien et al., 2003), the Organellar Genome Retrieval (OGRe) database (Jameson et al., 2003), the Joint Genome Institute (JGI/DOE) Organelle database (http://evogen.jgi.doe.gov/top_level/organelles.html), MamMiBase (Vasconcelos et al., 2005) and the NCBI Organelle Resources database (Wolfsberg et al., 2001), are specialized sources focused in organellar genomes that offer specific tools for the retrieval of detailed information related to complete mitochondrial genomes. Organellar databases usually offer more refined search engines than primary public repositories. These initiatives provide major contributions to comparative mitochondrial genomics. In this context, the Arthropod Mitochondrial Genomes Accessible database or AMiGA aims to provide the primary tools needed to optimize the access to sequence information and comparative analyses in order to stimulate the study of arthropod mitochondrial genomics.

2 STRUCTURE AND OPERATION

2.1 Infrastructure

AMiGA is implemented in a relational MySQL system in a GNU/Linux server (the organization of the data is shown in Table 1), and its web interface is based on an Apache 1.3 server with dynamic content generated by Perl CGI scripts, including the Bioperl module (Stajich et al., 2002). AMiGA contains sequences and related information from GenBank and NCBI-Taxonomy (http://www.ncbi.nih.gov/Taxonomy/) and runs automatic daily searches for new entries and updated information of current mitochondrial genomes. When a new mitochondrial genome is included in the NCBI RefSeq collection (Pruitt et al., 2005), AMiGA automatically substitutes its corresponding non-RefSeq file with this new one, which has been curated. Then, the new genome passes by an AMiGA curation process to uniformize nomenclature and, if necessary, AMiGA notes are included to provide additional or more accurate information about the genome. This updating system provides AMiGA with the most recent information available for complete arthropodan mitochondrial genomes.

2.2 Operation

Genome searches in AMiGA can be done using keywords for a taxon, a species’ common name, the GenBank accession number, author or journal names, or year of submission, with logical operators such as AND, OR and NOT. The search can also be done using an alphabetically ordered list of species or by taxonomical level. This flexibility in searching is determined by the way the information is stored in the database (Table 1). The results are sorted by the NCBI submission date, by the most recently updated NCBI files or by alphabetical order. Following the selection of an entry from a diversified set of keywords and additional gene choices, the user has

*To whom correspondence should be addressed.
Table 1. AMiGA data organization

<table>
<thead>
<tr>
<th>Category</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accession</td>
<td>Accession, is RefSeq, accession of non-RefSeq sequence</td>
</tr>
<tr>
<td>Sequence</td>
<td>Sequence, product, notes, sequence strand</td>
</tr>
<tr>
<td>Sequence type</td>
<td>Sequence type</td>
</tr>
<tr>
<td>Range</td>
<td>Sequence size, begin, end</td>
</tr>
<tr>
<td>Reference</td>
<td>Authors, GenBank submission date, institution, address</td>
</tr>
<tr>
<td>Submission</td>
<td>Authors, title, journal, PubMed, publication year</td>
</tr>
<tr>
<td>Species</td>
<td>Species name, common name, further information</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>Taxon, taxon hierarchy, Taxon NCBI ID</td>
</tr>
<tr>
<td>AMiGA Notes</td>
<td>note, field, comments</td>
</tr>
</tbody>
</table>

five options of results: detailed information, FASTA files, genomic graphics, multiple alignments and primer annealing sites.

(1) Detailed information. The results of this search are displayed in a tree structure with a folder for each genome and sub-folders with relevant information about bibliographic references, nucleotide and amino acid sequences, codon usage tables, and taxonomy and genome graphical views.

(2) FASTA files. The user can download the full genome sequence, individual genes or the main non-coding region, for a diversified set of organisms. AMiGA generates one FASTA file containing all of the selected sequences and one file per sequence, making it possible to analyze several genes at once.

(3) Graphics. Two types of graphics, linear and circular, are available through this menu. These graphics are the same as the detailed information menu, except that they are displayed for all of the selected species at once on the same page. For the linear graphics, the user can choose which features to show among protein genes, RNAs, overlaps, miscellaneous features and intergenic regions.

(4) Alignment. A major component of the AMiGA database is an online, multiple global alignment tool for nucleotide and amino acid sequences for one or multiple genes using the software ClustalW (Thompson et al., 1994). When selecting multiple gene analysis, an alignment for each gene is provided, in addition to the concatenated output retrieved from these individual alignments.

(5) Primer Finder. AMiGA uses software PrimerMatch (www.cbcb.umd.edu) to find primer annealing sites throughout the selected genomes. Degenerate primers can be entered using IUPAC ambiguity codes. The potential annealing sites are shown over the genome graphics.

2.3 Tools
AMiGA also provides a set of useful tools such as

- a chromatogram processing tool with a graphic interface based on phred (Ewing et al., 1998),
- a page with all recent AMiGA updates and inclusions, also accessible via RSS feed.

2.4 Password protected area
The AMiGA structure allows any user to create a password-protected area in which unpublished annotated genomes, complete or partial, can be submitted for analysis and comparison with the other genomes available in the database. A JavaScript script included in registration page generates the password’s MD5 digest encryption at the submission to increase security.

3 FUTURE DEVELOPMENTS
A broader database for metazoan mitochondrial genomes, MetAMiGA, has been built based on the original AMiGA platform and is available for testing on the AMiGA main page. AMiGA should contribute to our knowledge of arthropodan mitochondrial genomics and stimulate further research on organismal and organellar diversity and evolution.

ACKNOWLEDGEMENTS
We thank Renato Vicentini dos Santos for helpful advice. This work and L.S.N. and A.C.I. were supported by CNPq (PROFIX, grant no. 540.602/2001-9). P.C.F and A.M.L.A.E. were supported by FAPESP (grants no. 03/01458-9 and 04/09654-4).

Conflict of Interest: none declared.

REFERENCES